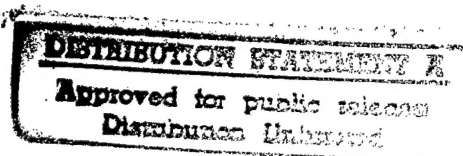


# En Route Generic Airspace Evaluation

Jerry A. Guttman, PERI  
Earl S. Stein, Ph.D., ACT-530



December 1997

DOT/FAA/CT-TN97/7

Document is available to the public  
through the National Technical Information  
Service, Springfield, Virginia 22161

U.S. Department of Transportation  
Federal Aviation Administration  
William J. Hughes Technical Center  
Atlantic City International Airport, NJ 08405

DTIC QUALITY INSPECTED 3

19980406 005

## NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof.

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

1. Report No. DOT/FAA/CT-TN97/7		2. Government Accession No.		3. Recipient's Catalog No.	
4. Title and Subtitle  En Route Generic Airspace Evaluation				5. Report Date December 1997	
				6. Performing Organization Code ACT-530	
7. Author(s) Jerry A. Guttman, PERI and Earl S. Stein, Ph.D., ACT-530				8. Performing Organization Report No. DOT/FAA/CT-TN97/7	
9. Performing Organization Name and Address Federal Aviation Administration William J. Hughes Technical Center Atlantic City International Airport, NJ 08405				10. Work Unit No. (TRAIS)	
				11. Contract or Grant No. DTFA03-93-C-00032	
12. Sponsoring Agency Name and Address Federal Aviation Administration Chief Scientist for Human Factors 800 Independence Ave., S.W. Washington, DC 20591				13. Type of Report and Period Covered  Technical Note	
				14. Sponsoring Agency Code AAR-100	
15. Supplementary Notes					
<p>16. Abstract</p> <p>This En Route Generic Airspace Evaluation is one of a series of air traffic control (ATC) simulation experiments. It is directed toward development and validation of the use of generic airspace for use in ATC research and development. For this project, generic refers to a sector that embodies the important elements of an en route sector including airways, en route radar performance, restricted areas, and radar procedures. In a generic sector, conditions are standardized. This is a significant advantage over using each controller's home sector where many factors vary such as familiarity and sector complexity. Experienced Federal Aviation Administration personnel developed and tested this en route generic airspace. The design was based on a typical high-altitude sector used in many en route centers. In addition, the sector was designed to facilitate rapid learning. In this experiment, experienced controllers performed their normal functions working with realistic traffic scenarios presented by a high fidelity ATC simulator. Voice communication equipment enabled controllers to issue commands to remote simulation pilots. The results showed that three performance-measurement categories (Air Traffic Workload Input Technique (ATWIT) ratings, system effectiveness measures, and controller self ratings of performance) showed high correlations between the generic and home sectors.</p>					
17. Key Words En Route Airspace Generic Sector ATC Performance Measurement ATC Performance Standards				18. Distribution Statement  This document is available to the public through the National Technical Information Service, Springfield, Virginia, 22161	
19. Security Classif. (of this report) Unclassified		20. Security Classif. (of this page) Unclassified		21. No. of Pages 63	
22. Price					

## Acknowledgments

We wish to acknowledge the contributions of Greg Bing (Jacksonville Center) for development of the generic sector, experimental briefings, and test support. We are grateful to Kathy Mann, Stan Gromelski, and Paul Stringer (PERI) for test support. This research would not have been possible without the technical support of Dennis Filler, Mike Pomykacz, Albert Macias, Mary Delemarre (ACT-510), and George Rowand (SRC).





## Table of Contents

	Page
Acknowledgments .....	iii
Executive Summary .....	vii
1. Introduction .....	1
1.1 Problem Statement .....	1
1.2 Assumptions and Goals .....	2
1.3 Review of the Related Literature .....	2
2. Experiment .....	9
2.1 Purpose .....	9
2.2 Logic Behind a Generic Sector .....	9
2.3 Airspace and Traffic Scenarios .....	10
2.3.1 Generic Sector Airspace and Scenarios .....	10
2.3.2 Home Center Airspace and Scenarios .....	13
3. Method .....	15
3.1 Participants .....	15
3.2 Simulation Facility .....	15
3.3 Experimental Design .....	15
3.4 Procedure .....	18
4. Results .....	19
4.1 Overview .....	19
4.2 Practice and Learning Effects Associated with the Generic Sector .....	19
4.2.1 Means and Standard Deviations for Dependent Measures .....	19
4.2.2 Orthogonal Components Analysis for Dependent Measures .....	22
4.3 Correlational Analyses for Generic and Home Sector Performance Scores .....	25
4.3.1 Reliability Analyses for Dependent Measures .....	25
4.3.2 Correlational Relationships Between Generic and Home Sector Performance Scores .....	28
4.3.3 Correlational Analyses Between ATWIT Ratings and Controller Self Ratings of Performance .....	31
4.4 Final Questionnaire Comments on the Entire Experiment .....	32
5. Summary and Conclusions .....	33
5.1 Discussion of Learning Rate for the Generic Sector .....	34
5.2 Discussion of Correlational Relationships Between Performance Scores .....	34
5.2.1 Discussion of Reliability of Performance Scores .....	34
5.2.2 Discussion of Correlations Between Performance Scores on the Home Sector and Generic Sector .....	35
References .....	37

## Table of Contents (Continued)

### Appendixes

- A - Demographic Form
- B - Observer Evaluation Form
- C - Transcript of Controller Final Questionnaire Comments

### List of Illustrations

Figures	Page
1. Generic Sector Radar Map .....	11
2. Adjacent Facilities and Their Radio Frequencies.....	12
3. Simulated Sector .....	14
Tables	Page
1. A Summary of the Experimental Design .....	16
2. The Presentation Order of Scenarios and Counterbalancing Features of the Experimental Design.....	17
3. Means and Standard Deviations for Blocks of Generic Sector Trials for Selected System Effectiveness Variables .....	20
4. Means and Standard Deviations for Blocks of Generic Sector Trials for Over-the-Shoulder Ratings.....	21
5. Means and Standard Deviations for Blocks of Generic Sector Trials for Controller Self Ratings.....	22
6. Orthogonal Components Analysis for Selected System Effectiveness Variables .....	23
7. Orthogonal Components Analysis for Over-the-Shoulder Ratings.....	24
8. Orthogonal Components Analysis for Controller Self Ratings.....	25
9. Reliability Analysis for Selected System Effectiveness Variables.....	26
10. Reliability Analysis for Over-the-Shoulder Ratings .....	27
11. Reliability Analysis For Controller Self Ratings.....	28
12. Correlation Between Home Sector and Generic Sector Blocks for Selected System Effectiveness Variables .....	29
13. Correlation Between Home Sector and Generic Sector Blocks for Over-the-Shoulder Ratings.....	30
14. Correlation Between Home Sector and Generic Sector Blocks for Controller Self Ratings.....	31
15. Correlation Between Average ATWIT Scores and Controller Self Ratings of Performance.....	32
16. Summary of Controller Final Questionnaire Comments.....	33

## Executive Summary

The Federal Aviation Administration (FAA) is constantly working to improve safety and system capacity. One approach to safety improvement is to enhance Air Traffic Control Specialist (ATCS) performance and reduce the probability of operational errors. The keys to improved performance include advances in personnel selection methods, training, and equipment design. The relationship between system effectiveness, safety and capacity, and controller performance, however, is more complicated and difficult to measure than the number of errors alone predict. The challenge for researchers is to establish methods to measure human performance in ways that directly relate to system effectiveness. This En Route Generic Airspace Evaluation is one of a series of air traffic control (ATC) simulation experiments directed toward development and validation of a reliable set of controller performance and system effectiveness measurement tools.

There is a problem when controllers come from different facilities or areas to participate in performance evaluations. All controllers know their home sector or area best. However, their performance may vary depending on the amount of time they have been working on the sector. In addition, sectors vary in their complexity and in degree of difficulty (Mogford, Murphy, Yastrop, Guttman, & Roske-Hofstrand, 1993). In a generic sector, conditions are standardized. This is a significant advantage over using performance measured on each controller's home sector where factors such as familiarity and sector complexity vary.

This research evaluated the feasibility of using airspace models that the participating controllers have not seen before and have not overlearned with practice. The use of generic airspace can simplify and reduce the cost of training and selection if personnel are able to perform relatively as well as they can with an over-learned environment.

Eighteen air traffic controllers from an Air Route Traffic Control Center participated in the study at the William J. Hughes Technical Center Research Development and Human Factors Laboratory at Atlantic City International Airport, New Jersey. The experimental apparatus consisted of a high fidelity ATC simulator with voice communication equipment to allow controllers to issue commands to remote simulation pilots. Each controller performed 11 different scenarios over 3 days of testing. The first 2 days of testing involved training for controllers who performed one scenario on the home sector followed by six runs on the generic sector. All traffic runs were of moderately busy traffic volume. The third day of testing was a test day where controllers performed four 1-hour runs. Two of these were on the home sector and two were on the generic sector. Traffic runs consisted of approximately nine aircraft every 15 minutes.

Experimenters collected data on ATCS performance, workload, system effectiveness, and self-assessment during the simulation. System effectiveness measures included the number of controller transmissions, number of altitude changes, and traffic. The Air Traffic Workload Input Technique (ATWIT) consisted of participants rating their workload level as they controlled traffic, and several questionnaires captured subjective ratings from participants. A demographic questionnaire requested background information from each participant. After each scenario, controllers made self assessment ratings of their own performance in a post-scenario

questionnaire. A final questionnaire at the end of the simulation measured subjective impressions of the realism of the simulation and the representativeness of the generic sector.

Three of four performance categories showed high and consistent correlations between the generic and home sectors. These categories were ATWIT ratings, system effectiveness measures, and controller self ratings of performance. These correlations suggest that controller workload, communication, and task management were basically the same, regardless of the sector configuration. Workload, as measured by ATWIT, was also highly correlated between the home sector and the fourth block of generic runs. This result suggests that once the controllers learned the sector, the workload was basically the same, regardless of the sector configuration. The results also indicated that system performance, as measured by system effectiveness measures, was very similar in both sector configurations. Fourteen of the 18 controllers thought that the generic sector was representative of a typical sector.

## 1. Introduction

### 1.1 Problem Statement

In 1994, there were 772 controller operational errors in the United States (FAA, 1996). This represented a slight increase from the previous year. The Federal Aviation Administration (FAA) is constantly working to reduce the probability of these errors. The keys to reducing controller errors involve selection, training, and equipage. Performance, however, is more complicated than the nature and volume of errors alone would predict. Researchers must define human performance in situation-specific contexts and establish methods to precisely measure it.

Performance has many definitions. For example, in his book, *Human Performance Engineering: A Guide for System Designers*, Bailey (1982) stated: "Performance then is defined as the result of a pattern of actions carried out to satisfy an objective according to some standard. These actions may include observable behavior or non observable intellectual processing (e.g., problem solving, decision making, planning and reasoning). Things change when people perform" (p. 4). "People working in different systems do share the common dimension of being somebody, doing something, someplace" (p. 1).

In this research, the operational definition of performance is the accomplishment of a task or interrelated set of tasks in relation to a defined and specified standard while operating within constraints of space, time, and resources. The concept of performance implies the ability to vary along a continuum of quality based on a wide variety of variables.

A human operator is part of this system. The operator must accomplish something in relation to a specified standard. Behavior is successful if it includes safe and expeditious airspace control that meets the current standard. The distance above or below the standard determines different levels of accomplishment within the unsuccessful and successful categories, respectively.

Human performance is situation specific, particularly in air traffic control (ATC). Situational variables include unique airspace, terrain below the airspace, weather conditions, and adjacent facilities and the agreements established with them. Facilities vary on how they emphasize operational concepts, which can influence human performance.

These variations complicate the task of developing generalizable performance concepts. They add a dimension to issues that relate to selection and training of controllers. Researchers raise the question: Can they design an airspace model that will generalize across the common dimensions of ATC as practiced in many facilities? Could such a model be easy to learn and would experienced personnel perform in it similarly to the way they perform in their home station airspace? If such airspace works, we may be able to use it for training. The process that leads to the creation of the airspace may also enhance our understanding of effective controller performance in ATC. These issues are the essence of the problem for this current research.

## 1.2 Assumptions and Goals

People learn new skills in different ways. Some skills are based on an absolute standard of performance that anyone in the trade can clearly define and easily recognize. In contrast, a relative standard is one that assumes explicitly or implicitly that there are many ways of looking at and evaluating performance. Trainers' understanding of what it takes to perform the task set is based on their own experience, structures, and standards (Berlinger, Angell, & Shearer, 1964). In this situation, the training system is very much dependent on the trainer and how all the other trainees are doing. Relative standards make performance measurement very complicated.

In ATC, there are some absolute or minimum standards against which the system judges everyone. The minimum separation allowed between aircraft under positive radar control is one of the most fundamental standards. This is an absolute standard, and everyone in ATC meets it or risks being removed. This means that this minimum standard is not very useful for looking at the range of performance that controllers, like all human operators, produce.

The airspace system has evolved with relative standards by using an over-the-shoulder rating scale as the basic metric. This is open to considerable latitude in interpretation (FAA, 1990). Evaluators apply their experience and biases when doing a controller evaluation. Fortunately, in a research environment, there are evaluation tools that may not be available in a field setting.

The foundation for this research is the assumption that performance of air traffic controllers can be measured in a number of ways. The quality of this measurement can continually improve, and this improvement is a worthwhile endeavor. These are basic assumptions. To the extent that behavior exceeds the current standard, observers will evaluate it as successful. If the behavior fails to meet the standard, they will view it as unsuccessful. The distance above or below the standard determines different levels of accomplishment within the unsuccessful and successful categories, respectively. Past experience has demonstrated that a range of acceptable behavior exists in all complex command and control systems and that simulation can be effective in stimulating this behavioral range. Given these assumptions, this program has a number of goals.

The FAA William J. Hughes Technical Center conducted this research for several reasons. The work is one element in an overall program on controller performance and error reduction. The research will evaluate the feasibility of using airspace models for testing and training. The use of generic airspace can simplify and reduce the cost of training and selection if personnel are able to perform relatively as well with it as they can with a well-known environment.

This study is the second in a series of research efforts done at the William J. Hughes Technical Center. Guttman, Stein, and Gromelski (1995) completed the first study, which focused on terminal operations.

## 1.3 Review of the Related Literature

Thorndike (1982) stated: "It is difficult even to formulate any complete definition of success on the job, much less develop a measure that adequately represents it" (p. 193). Most performance indicators are partial and incomplete. According to Thorndike, they lack range and time span.

They only provide a snapshot at best. Irrelevant sources of variance such as rater biases and low or unknown reliability can confound criteria. There are relatively few jobs for which a performance test is appropriate. It is necessary to determine what behaviors best represent the skill or what aspects of a product should be evaluated to determine performance. Thorndike concluded: "Performance evaluation (in many settings) tends to be subjective and unreliable at best" (p. 49).

Controller performance measurements have consistently involved tasks and variables derived from ATC and have produced findings expressed in ATC terms (Hopkin, 1980). Hopkin believed that it was also important to use basic psychological knowledge to explain and measure controller behavior. He felt researchers must consider the human side of ATC. Hopkin (1991) inferred that we may have to expand the more traditional view of performance to encompass concepts that we have dismissed in the past. There are obviously many different views related to the measurement of human behavior.

In an early comprehensive study of controller errors, Kinney, Spahn, and Amato (1977) analyzed FAA reports and developed categories of errors. These included: controlling in another's airspace, timing and completeness of flight data handling, inter-positional coordination of data, use of altitude on the display, procedures for scanning and observing flight data, phraseology and use of voice communications, and the use of human memory to include automatic capabilities. Kinney and his colleagues at MITRE Corporation spent considerable time in ATC facilities observing and talking with controllers. The error classification system they developed carried considerable weight for a number of years.

Based in part of the work of Kinney et al. (1977), the FAA decided to use a different set of categories to classify operational errors. Researchers classified operational errors for 1987 into the following categories: radar display, communication, coordination, aircraft observation, data posting, and position relief (FAA, 1988). By far, the most frequent source of errors identified by the FAA was in a subclass of radar display: the misuse of data. This category suggests that information was available, but operators either misinterpreted or inaccurately stored it in working memory. This overlaps several of the Kinney categories cited previously. Researchers, however, often use error rates and other error-related data as criterion variables without breaking the information into specific categories.

Rodgers (1993), for example, has associated controller error rates with the proportion of full performance level (FPL) controllers (those most practiced and proficient) assigned to an organization. Rodgers accomplished an analysis of the FAA operational error data base. He found that facility error rates were inversely proportional to the percentage of the work force that had achieved FPL status. Both the research community and operational management have used errors as performance indicators. For evaluating new systems or personnel who have already achieved FPL status, operational errors are a crude metric. However, they are metrics that have face validity for the ATC community. Researchers have continued to try to find a practical way of analyzing errors in field settings, and one group at the Civil Aeromedical Institute (CAMI) has succeeded.



Personnel at CAMI in Oklahoma City developed the Situation Assessment through Recreation of Incidents (SATORI) technique (Rodgers & Duke, 1993). SATORI analyzes system analysis report (SAR) tapes that contain all of the operational events for one radar position over a given time. Air route traffic control centers (ARTCCs) routinely record these tapes. The original purpose of SATORI was to evaluate the factors that led to an airspace incident or controller operational error.

Rodgers, Manning, and Kerr (1994) have taken the SATORI project one step further. They have developed the Performance and Objective Workload Research (POWER) program. This software package analyzes and computes many performance measures as they are described by Stein and Buckley (1992). These measurement tools focus on controller behavior in a naturalistic setting and emphasize the physical performance aspects of ATC. However, all behavior has both physical and psychological components. Thinking, also referred to as cognitive processing, is a critical area.

A group of researchers performed a cognitive task analysis of expertise to see if experts and novices differed in how they think (Seamster, Redding, Canon, Ryder, & Purcell, 1993). This represented an alternative view of controller performance. These researchers concluded that experts took a wider view of the evolving air traffic situation. Experts appeared to be more flexible in their approach to the dynamics in their airspace. The researchers identified en route controller tasks linked to their cognitive models of the airspace. These were: maintain situation awareness (SA), develop and revise the sector control plan, resolve aircraft conflicts, reroute aircraft, manage arrivals, manage departures, manage overflights, receive hand-offs, receive point-outs, initiate hand-offs, initiate point-outs, and issue advisories and safety alerts. Researchers have broken each of these into numerous subgoals. These establish the matrix of the controller's mental model.

According to Seamster et al. (1993), their research supports the hypothesis that experienced controllers group or organize their "picture" by events rather than by individual aircraft. The mental model and task accomplishment interact and influence each other. When thinking out complex ATC problems, experts used fewer but more varied planning strategies. The experts also had more strategies for managing their workload.

Endsley and Rodgers (1994) also focused on the cognitive aspects of controller performance. They studied en route ATC from the viewpoint of the information requirements for SA. The researchers attempted to identify the essential components of information that en route controllers must have in SA to perform their tasks. Using a panel of eight subject matter experts (SMEs), the researchers replayed ATC incidents to cue participant memory. The products of this work were a series of information requirements linked to each aspect of the controller's duties. This has implications for future performance evaluation. These elements of information may or may not appear in actual performance. A controller who does not acquire the critical elements of information may not perform as well as one who does.

How controllers think and use information has elicited considerable interest and research. Much of the work reviewed to this point has included theoretical formulations based on data already

available coupled with subject matter expertise. Researchers can employ active simulation to evaluate performance under controlled conditions.

In a study of SA, Endsley and Kiris (1995) applied a simulated driving task to examine the potential impact of automation on performance. They expressed concern about the potential loss of manual skills and awareness of the state of the system. They used a computer simulation of automobile navigation with a series of automated supports. Decision response time was the primary dependent variable. The hypothesis was that increased automation produces increased response latencies. The results supported this hypothesis. Decisions and the implied information processing behind them were longer. This occurred even after the researchers shut down the automation according to the research plan. The primary impact of automation was in the time dimension. The authors noted that although it took longer, participants eventually and usually made the correct decisions. The authors did not attempt to generalize these results to the more complex world of ATC. Although they did carefully define their terms in this study, they also tended to treat SA as a causative factor rather than an intervening variable.

Flach (1995) expressed concern about the construct of SA in the performance literature. He cautioned readers against the assumption that SA is a form of performance. He noted that researchers cannot measure SA directly and must infer it based on other behavior or errors. Flach suggested that SA has two important characteristics that serve performance research. First, it promotes the importance of good laboratory analogs to the real world because poor models and simulations will not create the appropriate internal psychological states and, therefore, could not be generalized. Flach also stated: "The test of the SA construct will be its ability to be defined in terms of objective, clearly specified, independent, and dependent variables" (p. 154). Simulation is one way to create laboratory models in which investigators can specify and manipulate variables.

Laboratories have used simulation research to study ATC concepts, equipment, and procedures for 35 years or more. Over this period, various sets of dependent variables have evolved to assist in the evaluation of system and individual controller performance. Research goals have involved tailoring the specific subset of variables to meet the needs of each study. The William J. Hughes Technical Center has conducted most of the ATC simulation studies.

In this research, a basic assumption is that everything that occurs in the simulation is recordable and recoverable on a post hoc basis. There is a data flow from target generation through controller actions and subsequent results. This occurs because aircraft responses and the relationships of all aircraft in the simulated airspace will be recoverable on a post-simulation basis. All raw data, such as the relative position of aircraft, are saved so that researchers can accomplish additional analyses as desired.

Researchers in ATC performance generally must establish the measures that they use. Stein and Buckley (1992) assembled and consolidated the variables that had been useful over the years for researchers at the FAA William J. Hughes Technical Center. The authors based this work on the research of Buckley, DeBaryshe, Hitchner and Kohn (1983) and Stein (1984a, 1984b, 1985).

Frequencies of events and time are the most widely used measures of the dependent variables. These may be discrete or cumulative and are based on a specific period. Research design, for example, can include a hypothesis of change in conflict frequencies and time duration based on the amount of time that a controller has been on position. So it is important to have the capability to compute statistics on predetermined time blocks.

Researchers have used these frequency performance measures in numerous studies over the years to evaluate concepts and systems. However, it has been argued, with some justification, that researchers can not always clearly define the difference between systems and individual performance measures. The two often mesh. Following is a description of a subset of controller performance research studies done at the Technical Center.

Buckley, O'Connor, Beebe, Adams and MacDonald (1969) conducted a simulation study of air traffic controllers. These researchers focused on the assessment of controller performance and its relationship to chronological age. Buckley and his colleagues used a combination of objective system measures and over-the-shoulder SME ratings. They commented that "a difficulty with such subjective ratings is their frequent unreliability" (p. 49). They employed eight observers in 2-person teams who did over-the-shoulder ratings. The observers were current controllers from facilities other than those where the participants worked. The correlations between pairs of raters ranged from .06 to .72, using intraclass correlations as the indicator of inter-rater reliability. There was a considerable range of reliability coefficients and preponderance of low relationships. Buckley et al. moved on to even more extensive performance research.

Buckley et al. (1983) performed two experiments to examine the use of simulation for performance evaluation. They emphasized the quality of measurement and identified the basic dimensions for measuring ATC functions in real time. They studied the interaction of sector geometry and density. There were also statistically significant simple effects of sector geometry and traffic density for almost all of the 10 performance measures. The authors suggested that "the nature and extent of this interaction depends on the measures involved" (p. 73). The fact that sector geometry influenced performance, as measured, is an ongoing concern when dealing with the possibility of generic airspace sectors.

A second experiment involved collecting a great deal of data over time by repeated measures. The data base was sufficient so that researchers could compute a factor analysis to look for redundancy in the measures used to quantify system performance. Each of 39 controllers participated in 12 one-hour runs using the same sector with the same traffic level.

The data resulting from the first Buckley et al. (1983) experiment were cross-validated with the factor analysis derived from the second experiment. This produced four meaningful factors or measures: confliction, occupancy, communication, and delay. The confliction factor had measures of 3-, 4-, and 5-mile conflicts. The occupancy factor contained measures of the time an aircraft was under control, distance flown under control, fuel consumption under control, and time within boundary. The communications factor involved path changes, number of ground-to-air communications, and duration of ground-to-air communications. The delay factor included total number of delays and total delay times. Two auxiliary measures, number of aircraft handled

and fuel consumption, were also relevant. These experiments conducted by Buckley et al. have served as building blocks for most of the controller performance research that has followed.

Using a subset of the Buckley et al. (1983) measures, another study compared parallel approach separation standards between 1.5 and 2 nmi. Variables included not only controller operational errors but also many other data variables to include the landing rates at the airport under study. Results demonstrated that controller performance did not decline and that there was no increase in subjective estimates of workload. The landing rates were higher for the reduced separation standard (Stein, 1989). However, researchers must always exercise caution when they complete a study on significant differences that might have occurred or were hypothesized but that do not materialize. If the result does not reflect actual lack of differences in the real world to which we would like to generalize, it is very difficult to calculate the probability of that error.

Simulation performance measurement has been and is being used in the William J. Hughes Technical Center Research Development and Human Factors Laboratory (RDHFL). To a certain extent, measurement fell into a pattern that stressed frequency and time variables. However, Paul (1989, 1990) created a unique tool for use in ATC simulation research, the Aircraft Proximity Index (API). This tool takes an entirely new look at conflicts between aircraft. Instead of simply counting them, the API provides a graded severity scale ranging from 0 to 100. As long as it is 0, there is no conflict and, as the numbers rise, so does the severity. An API of 100 is a score that means a collision is imminent. Instead of assuming that all conflicts are alike, this tool takes into consideration horizontal and vertical separation and the actual slant range distance between aircraft. Research personnel now routinely use the API in ATC simulations at the Technical Center.

Sollenberger and Stein (1995) used all the measures then available including those by Buckley et al. (1983) and Paul (1990). They conducted a study of controller memory issues to determine whether they could enhance performance using a memory aid. Sixteen controllers worked in simulated TRACON airspace. Researchers evaluated their performance using automated tools and over-the-shoulder observation.

The memory aides did have some positive influence on controllers' behavior, as recorded in the automated performance measurement data. In the aided condition, controllers made significantly fewer ground-to-air transmissions. Also, they gave fewer changes of altitude and heading. Researchers have used these variables as indicators of controller workload. Controllers made fewer hand-off errors when they had the memory aids as compared to when they did not have them. Without the wide range of performance indicators, researchers may not have correctly identified these differences.

Guttman et al. (1995) completed another study of controller performance under two different sets of airspace conditions. In one, the controllers were familiar with the airspace. The other was a generic terminal radar approach model that controllers had not used before. This study preceded the research reported here. It evaluated controller performance under both conditions to see if researchers and trainers could use the generic model for their respective needs. Researchers also wanted to evaluate generic sectors as tools for controller performance evaluation.

Performance indicators on many quantitative variables were similar across the two types of airspace. Controllers were able to learn the generic airspace rather quickly, and performance variables did not change appreciably over the course of familiarization with the generic sector. The use of automated data collection supported the conclusion that the sector was easy to learn and did not lead to performance decrements once some learning had occurred over a 3- to 4-hour period.

This study also used an over-the-shoulder observer who rated the performance of the participants and estimated how hard they were working. This type of behavioral observation and evaluation is difficult. It requires the ability of SMEs to accept training and forego long established biases. Subject matter expertise and knowledge are basic requirements for evaluating the performance of others. Someone without knowledge and experience would not know what to look for or how to apply any conventional standard of performance. However, when the official standard is not clearly defined, SMEs may continue to apply their own unique standards in place of the designated standard.

Experience, training, the performance of current peers, and, possibly, the organizational standards influence internal standards. It is possible that these mental models are more alike within a facility than they are across facilities. However, while there are few certainties when it comes to human performance, the following statement by Bailey (1982) is accurate, "... people do not perform consistently and available measurement devices are imperfect" (p. 554). Despite these admonitions and the difficulty in doing effective performance rating, such evaluations are very popular and continue in business, industry, and government. They have face validity for many decision makers even when they fail to meet basic criteria for reliability and criterion-related validity. The Technical Center has been developing more effective training and research tools for performance evaluation.

Using data from the Guttman et al. (1995) simulation research, another study examined performance evaluations by SMEs who observed video playbacks of the simulations run earlier (Sollenberger, Stein, & Gromelski, 1997). The purpose of this research was to evaluate the reliability of a new performance rating form for use in research and test validation. The researchers identified observable actions for use in making behaviorally based performance ratings. Twenty-four rating scales assessed different areas of the controller's domain.

Researchers presented video tapes of controllers from a previously recorded simulation study on a multi-screen projection system. Six supervisors from different air traffic facilities participated as observers/raters. After a week-long training program, the observers viewed and rated 20 one-hour video tapes. The results indicated that the inter-rater reliability ranged between  $r = .70$  to  $r = .90$  for most of the rating scales. A few scales had relatively low reliability due, possibly, to the difficulty in accurately detecting and evaluating the observable actions. This research centered on task-related issues in ATC.

This research served as a building block for the current study. Each experience with ATC simulation on performance measurement has helped to build the knowledge base necessary to create and evaluate generic airspace as a viable research tool.

## 2. Experiment

### 2.1 Purpose

The purpose of this research was to develop and validate the concept of using an en route generic sector to evaluate air traffic controller performance. The study had two major goals. Evaluating the controller's ability to learn a new sector in a relatively short amount of time was the first goal. Evaluating the similarity and differences in performance across the generic and home sectors was the second goal.

### 2.2 Logic Behind a Generic Sector

Air traffic controller operations involve many tasks that are difficult to observe and measure such as image recognition, planning, and decision making. Individual controller style also affects performance. As a result, the process of developing reliable performance measures requires analysis of a large volume of data from different controllers. Ideally, the researcher collects these measures while simulating air traffic in the controller's home sector. However, a controller's performance may vary depending on the amount of time he or she has been working the sector. In addition, sectors vary in complexity and, therefore, in difficulty for the controller (Mogford, Murphy, Yastrop, Guttman, & Roske-Hofstrand, 1993). A standard generic sector could be a potential solution in that all the conditions under which performance is measured are the same for all participants. This is a significant advantage over using performance measured on each controller's home sector where many factors, such as familiarity and sector complexity, vary.

To perform this study, the researchers defined and developed a generic sector. In the context of this research, generic refers to a sector that embodies the important elements of an en route sector (i.e., airways, en route radar performance, restricted areas, and radar procedures). To achieve the goals of the study, the researchers designed the generic sector to have the same type of elements, but these elements were sometimes quite different from the home sector. For example, the home sector had airways running north and south. The generic sector had approximately the same number of airways and route length but had them running in an east/west direction. The reason for incorporating differences in the generic sector is that these differences require some learning on the part of the controller.

However, making a sector completely different from the home sector can introduce a number of potential confounds into the experimental design. Major items that were comparable included sector size, the mixture of traffic, the number and altitudes of the restricted areas, and distance traversed through the sector on an airway. Major items that were different included the Letters-of-Agreement (LOAs), the direction of traffic flow, and the placement and orientation of sector boundaries. The researchers felt that this mixture of similarities and differences produced a comparable generic sector that still requires learning on the part of the controller.



## 2.3 Airspace and Traffic Scenarios

### 2.3.1 Generic Sector Airspace and Scenarios

One of the primary concerns of this effort was that the generic airspace appear realistic to an FPL controller yet could be learned with a minimal amount of training. To achieve this objective, an Air Traffic Control Specialist (ATCS) participated in the development of the airspace and traffic scenarios. The ATCS was a current FAA en route controller who had extensive experience working in the en route environment.

Researchers based the generic airspace in this study on a typical high altitude sector used in many en route centers. As mentioned previously, major elements were matched with the home sector for experimental purposes. Arrival and overflight aircraft originated from one of two airways to the west. Additional overflight aircraft were generated either from a north-south airway or from one of two airways to the east. These two eastern airways converged at a single intersection (MIDDLE), and aircraft traveling these routes often had to merge with departure aircraft climbing out of Midtown Airport. Figure 1 illustrates the layout of the en route generic sector.

To expedite learning the fixes, the three letter identifiers for VORTACs corresponded to their magnetic heading or their position relative to the center of the radar map. For example, the northwest VORTAC was called NWT, and the southwest VORTAC was called SWT. An intersection close to the middle of the map was named MIDDLE intersection. Intersections near the upper and lower boundaries of the sector were named UPPER and LOWER, respectively. Another significant feature was the naming of the airways. Because airways are really just "highways in the sky," the naming conventions for interstate highways were used to name the jetroutes. East-west jetroutes were even two digit numbers and increased in magnitude the farther north they got (i.e., J64, J70, J74). The north-south jetroute was an odd two-digit number (J75).

The ATC SME developed LOAs to provide the participant controllers with standardized hand-off procedures. The SME also created four adjacent sectors, Alpha, Bravo, Charlie, and Genera Low. Genera Low was the sector immediately below the generic sector (Genera High) and employed an altitude structure from the ground to FL 230. The Genera High sector was responsible for altitudes from FL 230 to FL 500. Alpha sector was north of the generic sector, Bravo sector was east of the generic sector, and Charlie sector was due south. Figure 2 illustrates the adjacent sectors and their radio frequencies.

The traffic mixture for the generic sector was based on actual flights through the home center. The home center provided SAR tapes containing the flight data. Personnel from the William J. Hughes Technical Center extracted flight plan data for approximately 200 flights through one sector. These data were formed into a database and used for flight plans for all home sector and





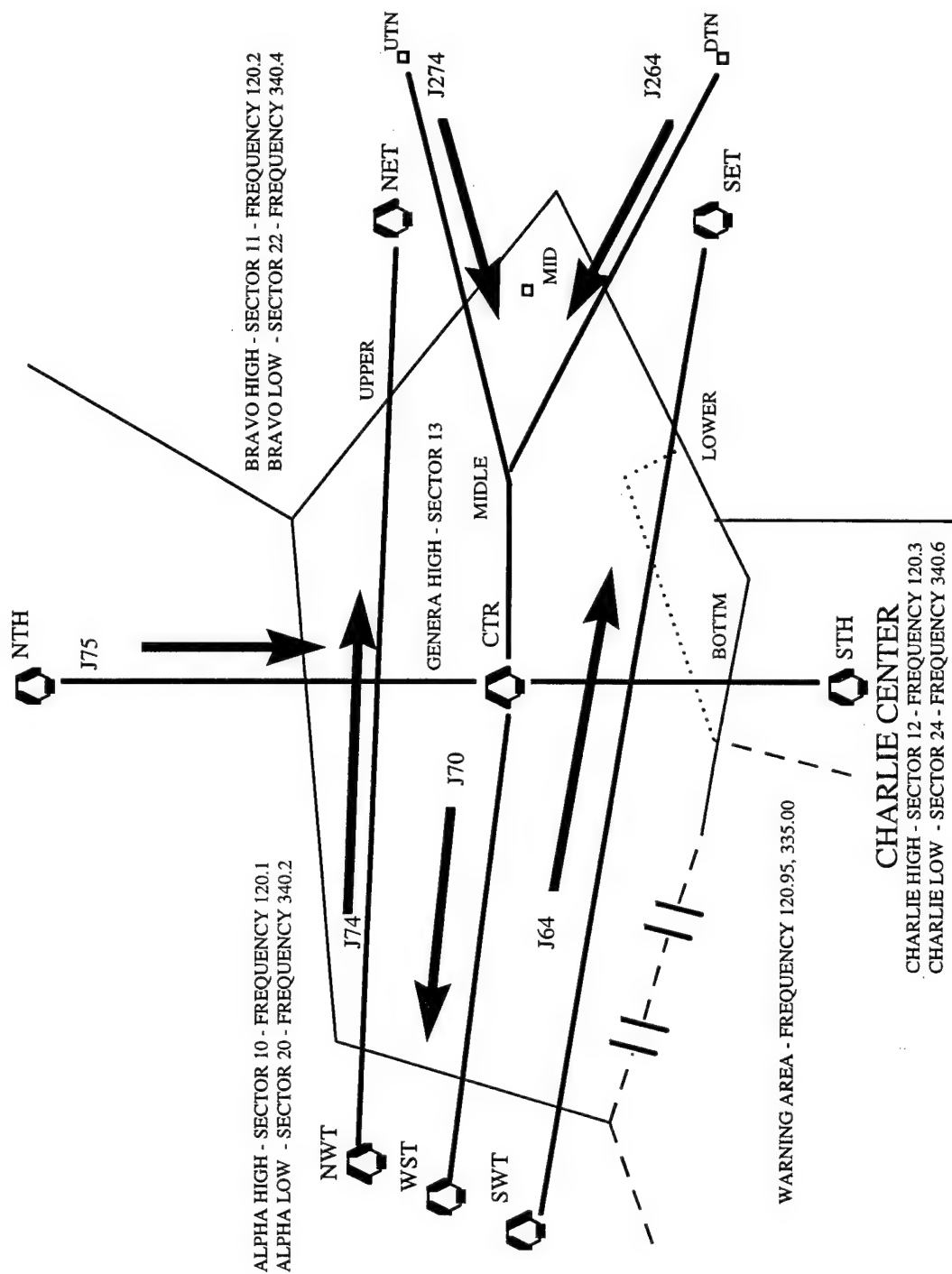


Figure 2. Adjacent facilities and their radio frequencies.

generic sector scenarios. Approximately 90% of the flight mixtures were air carriers flying medium to large transport aircraft (i.e., DC-9s, 727s, 737s, 747s, and L1011s). The remaining 10% of the mixture were general aviation aircraft including commuter jets (Learjet, Cessna Citation) and twin engine propeller driven aircraft (DeHavilland Dash 6).

Researchers constructed scenarios that accurately simulated traffic running through a typical high altitude sector. Traffic types included arrivals (descending traffic), departures (climbing traffic), and overflights. Arrival traffic was scheduled to land at one of two airports (Uptown via J74 or Downtown via J64). The controller's responsibility was to make sure these arriving flights were descended to FL 240 before leaving the sector boundaries. Departure traffic was generated from Midtown Airport. These target aircraft automatically climbed to FL 230 and leveled off. The controller's responsibility for these aircraft was to climb them to the requested altitude printed on the corresponding flight strip. Overflight traffic appeared on all airways, and the controller was responsible for safely merging this traffic with the departure and arrival aircraft. The controller also had to ensure that overflight traffic bound for the same airport had to maintain at least 10 nmi of lateral spacing. The scheduled rate of appearance of aircraft was representative of moderately busy traffic conditions.

### 2.3.2 Home Center Airspace and Scenarios

One of the primary concerns in this experiment was to create a realistic simulation of ARTCC airspace. Before the simulation, researchers gathered a large amount of data on the sector operations, normal operating procedures, and airspace boundaries. They used those data to create a realistic depiction of the home sector and construct realistic traffic scenarios. The researchers believed that the efforts invested in creating a realistic simulation of the home sector would motivate participants and increase the credibility of the research results. The research team constructed a radar map of the home airspace using the information obtained from the home center. Figure 3 illustrates this radar map.

The traffic mixture for the home sector was based on actual flight data recorded at the home center. The experimenters obtained SAR tapes that had flight plan information recorded on them. The researchers extracted flight plan data from the home sector from these SAR tapes and formed a database of that flight plan information. Most flights traveling through this sector are air carrier aircraft. The general aviation aircraft that do fly through the sector are small jets or twin engine commuter aircraft.

The experimenters reconstructed scenarios that accurately simulated traffic patterns in the home sector. Many of the aircraft call signs were familiar to controllers and represented common air carriers that operate in the home center. Flight types included arrival, departure, and overflight traffic. Arrival aircraft were scheduled to arrive at either Stuart or Vero Beach Airports. Controllers were responsible for descending these aircraft to FL 240 before leaving the sector boundaries. Departure aircraft were generated from Orlando International Airport or Orlando

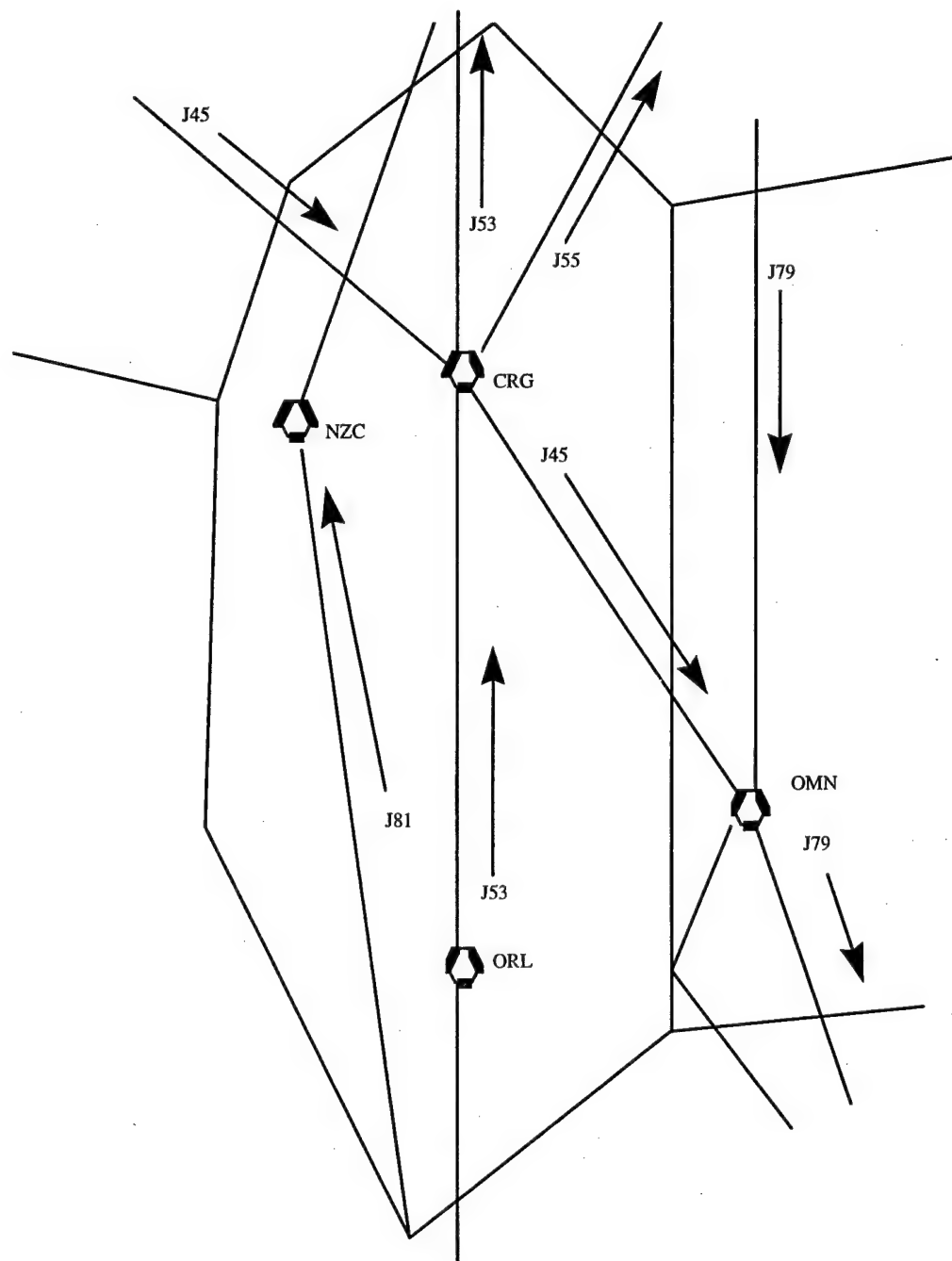


Figure 3. Simulated sector.

Executive Airport. These flights automatically climbed to FL 230 then leveled off. The controllers monitored aircraft climbing to the requested altitude printed on their flight progress strips. Overflight traffic was present on all the airways. The controller was responsible for merging and separating these flights from the departure and arrival aircraft. The scheduled rate of appearance for aircraft was set for moderately busy traffic conditions.

### 3. Method

#### 3.1 Participants

Eighteen air traffic controllers from an ARTCC volunteered for this study and researchers assured them of their anonymity and confidentiality. All participants were FPL controllers with normal or corrected-to-normal vision and had actively controlled traffic for the 12 months prior to the study. Each controller completed a Demographic Form describing their background characteristics. Controllers ranged in age from 29 to 52 years old (mean = 34.7,  $SD = 5.1$ ) and had 3 to 29 years of active service (mean = 8.5,  $SD = 6.0$ ). Controllers also provided self ratings for four personal attributes that could affect simulation performance on a scale ranging from 1 (low/poor) to 10 (high/good) on each question. A copy of the form is in Appendix A. The attributes included skill (mean = 8.6,  $SD = 1.1$ ), motivation (mean = 9.2,  $SD = 1.0$ ) and health (mean = 8.7  $SD = 1.0$ ). The final attribute was a measure of video game experience for hours-per-month (mean = 12.6,  $SD = 23.3$ ). Researchers have found that video game experience could have an impact on controller performance in a low fidelity simulation (Zingale, Gromelski, Ahmed, & Stein, 1993). However, such an effect was not anticipated in this high fidelity simulation study.

#### 3.2 Simulation Facility

The researchers conducted the experiment in the RDHFL at the William J. Hughes Technical Center at the Atlantic City International Airport, New Jersey. The experimental apparatus consisted of a state-of-the-art controller work station with a high resolution graphics display, voice communications equipment, networked computer resources, and ATCoach simulation software (copyright UFA Inc., 1995). A research psychologist and an ATCS who observed the participant in the experiment room and made over-the-shoulder ratings conducted the study. A voice communication link to another experiment room allowed the controller to issue clearances to personnel serving as simulation pilots. Two simulation pilots provided realistic voice feedback to the controller and controlled the movement of radar targets using keyboard commands. Additionally, the simulation pilots served as ghost controllers to simulate coordination with controllers in charge of adjacent sectors. As part of the simulation, flight progress strips for the entire scenario were printed and placed in a flight strip bay adjacent to the controller's work station. Controllers marked and arranged the flight strips as they were accustomed to doing in the ARTCC. The controllers previewed the flight strips before the start of the simulation to get a sense of the upcoming traffic situation. During the simulation, audio-visual equipment was used to record each participant's activities. Technicians videotaped the radar display and the controller as he or she controlled traffic during the simulation. They also recorded the audio from the simulation, which included controller and simulation pilot communications.

#### 3.3 Experimental Design

This was a quasi-experimental design. Quasi-experimental designs are often used in field research or a field setting where treatments differ on a number of variables, and experimental

control of a single variable is not possible (Gay, 1994). Such is the case when comparing or correlating performance on different sectors where many factors can differ.

Table 1 illustrates the experimental design for this study. The design follows a time series approach where a number of treatments are ordered chronologically and measurements are taken after each treatment. Each controller participated in 11 one-hour scenarios over a 3-day period. The first and second days were considered training days where the participant controlled traffic on a home sector scenario and then controlled traffic on six generic sector scenarios. These six generic sector scenarios were counterbalanced to evenly distribute any differences in difficulty that might exist.

Table 1. A Summary of the Experimental Design

DAY	SCENARIOS (1 Hour, Moderate Traffic Level)	
	Morning	Afternoon
First	1 Home	2 Generic <sup>a</sup>
Second	2 Generic <sup>a</sup>	2 Generic <sup>a</sup>
Third	2 Home	2 Generic

<sup>a</sup>counterbalanced

The four remaining runs were completed on a third-day session. This was a test day because, by this point, the participants had received nearly 2 days of hands-on training on the generic sector. Each controller worked two home sector scenarios in the morning and two generic sector scenarios in the afternoon.

For all scenarios, the traffic volume consisted of 37 aircraft generated in a 60-minute period. This corresponded to a rate of nearly 9 aircraft entering the scenario every 15 minutes. Each scenario included 11 departure flight plans, 5 arrival flight plans, and 21 overflight flight plans. The aircraft destinations and flight plans were not systematically ordered, so the traffic patterns were not predictable from working the previous scenario. However, the scenarios were matched for entry time of aircraft into the scenario. This was done to balance the flow of traffic into the scenario and the resulting taskload associated with working the traffic. Table 2 illustrates the presentation orders of scenarios and counterbalancing features of the experimental design.

The present experiment used a list of ATC performance measures that have been examined in previous research (Buckley et al., 1983; Stein & Buckley, 1992). The first category was system effectiveness. The current study focused on system effectiveness variables to include the number of conflicts, clustering of aircraft (complexity index), number of communications, number of clearances, and total distance the aircraft flew in the scenario. The second category of measures was controller workload, which was assessed through the Air Traffic Workload Input Technique (ATWIT) and through items on a post-scenario questionnaire. A third category was controller

**Table 2. The Presentation Order of Scenarios and Counterbalancing Features of the Experimental Design**

Subj.	DAY NUMBER 1			DAY NUMBER 2				DAY NUMBER 3			
1	Hm1 <sup>a</sup>	Gen1 <sup>b</sup>	Gen2 <sup>b</sup>	Gen3 <sup>b</sup>	Gen4 <sup>b</sup>	Gen5 <sup>b</sup>	Gen6 <sup>b</sup>	Hm3 <sup>a</sup>	Hm4 <sup>a</sup>	Gen7 <sup>b</sup>	Gen8 <sup>b</sup>
2	Hm2 <sup>a</sup>	Gen2 <sup>b</sup>	Gen3	Gen4	Gen5	Gen6	Gen1	Hm4 <sup>a</sup>	Hm3	Gen8	Gen7
3	Hm1	Gen3	Gen4	Gen5	Gen6	Gen1	Gen2	Hm3	Hm4	Gen7	Gen8
4	Hm2	Gen4	Gen5	Gen6	Gen1	Gen2	Gen3	Hm4	Hm3	Gen8	Gen7
5	Hm1	Gen5	Gen6	Gen1	Gen2	Gen3	Gen4	Hm3	Hm4	Gen7	Gen8
6	Hm2	Gen6	Gen1	Gen2	Gen3	Gen4	Gen5	Hm4	Hm3	Gen8	Gen7
7	Hm1	Gen1	Gen2	Gen3	Gen4	Gen5	Gen6	Hm3	Hm4	Gen7	Gen8
8	Hm2	Gen2	Gen3	Gen4	Gen5	Gen6	Gen1	Hm4	Hm3	Gen8	Gen7
9	Hm1	Gen3	Gen4	Gen5	Gen6	Gen1	Gen2	Hm3	Hm4	Gen7	Gen8
10	Hm2	Gen4	Gen5	Gen6	Gen1	Gen2	Gen3	Hm4	Hm3	Gen8	Gen7
11	Hm1	Gen5	Gen6	Gen1	Gen2	Gen3	Gen4	Hm3	Hm4	Gen7	Gen8
12	Hm2	Gen6	Gen1	Gen2	Gen3	Gen4	Gen5	Hm4	Hm3	Gen8	Gen7
13	Hm1	Gen1	Gen2	Gen3	Gen4	Gen5	Gen6	Hm3	Hm4	Gen7	Gen8
14	Hm2	Gen2	Gen3	Gen4	Gen5	Gen6	Gen1	Hm4	Hm3	Gen8	Gen7
15	Hm1	Gen3	Gen4	Gen5	Gen6	Gen1	Gen2	Hm3	Hm4	Gen7	Gen8
16	Hm2	Gen4	Gen5	Gen6	Gen1	Gen2	Gen3	Hm4	Hm3	Gen8	Gen7
17	Hm1	Gen5	Gen6	Gen1	Gen2	Gen3	Gen4	Hm3	Hm4	Gen7	Gen8
18	Hm2	Gen6	Gen1	Gen2	Gen3	Gen4	Gen5	Hm4	Hm3	Gen8	Gen7

<sup>a</sup>home sector scenarios

<sup>b</sup>generic sector scenarios

performance as measured by the Observer Evaluation Form found in Appendix B. It incorporated rating scales, which included some behavioral examples of what the scale was trying to measure. The observer rated on an eight-point scale. Twenty-four dimensions included the following areas: maintaining a safe and efficient traffic flow, maintaining attention and SA, prioritizing, providing control information, technical knowledge, and communicating. The remainder of the dimensions can be seen in Appendix B.

The controller's self assessment of his or her performance was the last measurement domain. A post-scenario questionnaire administered immediately after the controller finished the scenario measured this area (see Appendix C). The self-report ratings reflected categories used currently in en route centers for training and performance rating. Dimensions included communication, prioritization, safety, and technical knowledge. In addition, an item regarding the degree to

which the controller thought he or she could have improved with practice was added to examine the controller's self assessment of mastery on the generic sector.

### 3.4 Procedure

A training program assisted controllers in learning the generic sector and the procedures associated with controlling traffic in the sector. Researchers provided a training manual detailing the operating procedures and LOAs associated with the generic sector. This manual contained detailed maps of the sector layout and frequencies and names for the adjacent sectors. The participants had the manuals before they arrived for their first session.

When controllers arrived at the RDHFL, researchers briefed them on how the experiment was to be conducted, what was expected from them, and their rights as volunteers. At this point, the principal investigator asked each controller for their verbal informed consent to participate in the study. Next, an ATCS briefed each controller on the generic sector. This briefing included text presentations and visual aids, with a static presentation of the generic sector on the radar screen. The ATCS reviewed the LOAs, the fix names and locations, and direction of traffic for aircraft on the airways. The specialist also reviewed the slight differences that existed between the simulation and the operational software and hardware that existed in the field. One notable difference was the use of a software-generated computer readout device (CRD). This was generated in a window on the controller's workstation screen and controllers interacted with the soft buttons using the trackball. Each controller was given a chance to ask questions before working the first scenario.

On the first day session, each controller worked a home sector scenario to gain some experience using the simulator, interacting with the simulation pilots, and using the ATWIT device. The data from this first scenario were not used in any of the subsequent analyses. As controllers worked each scenario, an ATCS made over-the-shoulder observations of the controller's performance and completed the rating form. After each scenario, controllers completed a self-assessment of their own performance in a post-scenario questionnaire. At the conclusion of the final day of testing, researchers asked the participants to fill out a final questionnaire, giving them an opportunity to comment on their experiences.

Researchers measured controller workload in real time using ATWIT (Stein, 1985). ATWIT provides an unobtrusive and reliable means for collecting participants' ratings of workload as they control traffic. In the present study, a touch screen was used to present the workload rating scale and record the controller's responses. Controllers indicated their current workload by pressing one of the touch screen buttons labeled from 1 (very low) to 10 (very high). The device queried the controller every 5 minutes. The controller had 20 seconds to respond by touching one of the buttons. If they were too busy to respond within the 20 seconds, the maximum workload rating of 10 was recorded by default.

## 4. Results

### 4.1 Overview

The main results of this experiment appear in sections 4.2 and 4.3. Section 4.2 will present analyses collected during the day-1 and day-2 training sessions. These analyses focus on ratings for successive trials on the generic sector. They will examine the extent to which system effectiveness variables, workload ratings, and expert assessments of performance changed as controllers became more familiar with the generic sector. Section 4.3 analyzes correlational relationships within the generic sector runs to establish reliability with respect to system effectiveness, workload ratings, and expert assessments of performance. Section 4.3 also analyzes the correlations between performance scores on the generic and home sectors.

Section 4.4 will summarize the feedback that controllers provided about the experiment, and the results of the final questionnaire will be presented. The final questionnaire provided another means for evaluating the generic sector because many of the comments centered on how representative the generic sector was of the en route environment, the effectiveness of the training manual, the effectiveness of the hands-on training, and the realism of the simulation.

### 4.2 Practice and Learning Effects Associated with the Generic Sector

#### 4.2.1 Means and Standard Deviations for Dependent Measures

All trials for the following analyses are grouped into blocks of trials for ease of interpretation. Block 1 (B1) represents the average of the performance scores on trial 1 and trial 2. Block 2 (B2) represents the average of the performance scores on trial 3 and trial 4 and Block 3 (B3) represents the average of the performance scores on trial 5 and trial 6. A Block 4 (B4) was also included, which represents the average of the generic 7 and 8 scenarios.

In this portion of the experiment, the independent variable examined is practice, as presented by multiple blocks of trials. If there are significant differences representing improved performance between earlier and later trials with respect to dependent measures, the results suggest that learning occurred.

However, a lack of a significant result may have multiple interpretations, as the dependent measure may lack sensitivity to learning and more trials may be needed before a learning effect can be detected statistically. It also may mean that learning was not required. Table 3 presents a listing of means and standard deviations arranged by block for selected system effectiveness variables. In addition, statistical tests (analyses of variance [ANOVAs] and post hoc tests) were done to examine B2 vs. B1, B3 vs. B1, and B4 vs. B1. This was done to assess if any changes in learning occurred between earlier (B1) and later trials (B2, B3, and B4).

As shown in Table 3, most of the performance measures showed a high degree of stability in the earlier trials (B1 and B2). This significantly changed when compared to performance in the later trials (B3 and B4). This is based on the comparisons done between each block and the B1 data.



Table 3. Means and Standard Deviations for Blocks of Generic Sector Trials for Selected System Effectiveness Variables

VARIABLE	B1		B2		B3		B4	
	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
No. of Conflicts	0.19	(0.29)	0.09	(0.19)	0.14	(0.29)	0.03 <sup>a</sup>	(0.12)
No. of Special Conflicts	1.07	(0.41)	1.03	(0.44)	0.77	(0.35)	0.59 <sup>a</sup>	(0.45)
No. of Sector Conflicts	0.03	(0.12)	0.06	(0.16)	0.06	(0.16)	0.00	(0.00)
Complexity Index	1.13	(0.20)	1.09	(0.16)	1.04	(0.20)	0.46 <sup>a</sup>	(0.12)
No. of Altitude Changes	27.64	(2.70)	27.21	(2.11)	26.73	(3.28)	26.07	(2.85)
No. Heading Changes	9.43	(4.03)	10.35	(3.71)	10.35	(4.60)	5.54 <sup>a</sup>	(4.55)
No. of Speed Changes	6.42	(3.51)	7.42	(2.80)	7.13	(3.04)	9.52 <sup>a</sup>	(4.02)
PTT - Number	148.92	(19.5)	151.75	(20.0)	148.17	(16.0)	150.44	(15.0)
PTT - Duration	603.1	(106)	589.0	(98.3)	570.0	(79.9)	581.3	(95.1)
No. of Controller Key	106.17	(2.15)	72.69	(8.54)	74.98	(1.67)	75.33	(2.20)
No. of Sim-Pilot Key	268.36	(40.9)	277.06	(35.6)	272.69	(37.1)	250.17	(53.3)
No. of Aircraft Handled	37.87	(0.57)	37.82	(0.37)	38.06	(0.45)	37.94	(0.53)
Time Controlled (sec.)	25194	(1903)	25041	(5478)	23443	(2443)	21594 <sup>a</sup>	(2897)
Distance Flown (miles)	2831.9	(213)	2809.0	(614)	2634.5	(283)	2424.4 <sup>a</sup>	(327)
No. of Handoffs Accept	37.27	(0.64)	37.27	(0.59)	37.57	(0.69)	37.10	(0.62)
No. Handoffs Delayed	0.68	(1.67)	0.47	(1.39)	1.20	(2.63)	0.49	(1.51)
Perc. Flights Complete	0.88	(0.06)	0.91	(0.07)	0.93 <sup>a</sup>	(0.07)	0.93 <sup>a</sup>	(0.06)
Average PTT time (sec.)	4.05	(0.46)	3.89	(0.52)	3.86 <sup>a</sup>	(0.48)	3.87 <sup>a</sup>	(0.54)
PTT/Aircraft	3.93	(0.51)	4.01	(0.52)	3.90	(0.44)	3.97	(0.40)
Conflicts/Aircraft	0.0051	(0.01)	0.0024	(0.01)	0.0032	(0.01)	0.0015	(0.01)

<sup>a</sup>significantly different from Block 1 ( $p < .05$ ) as determined by the Tukey post hoc test.

This change was also in the expected direction for many of these variables. For example, conflicts per aircraft was reduced from 5 aircraft per thousand (B1) to 1.5 aircraft per thousand (B4). Average push-to-talk (PTT) time was significantly shorter for the B4 block of runs compared to B1. Distance flown and time under control was also decreased, suggesting a more efficient use of control techniques in the later generic runs. Complexity or clustering of aircraft was also significantly reduced in the B4 runs compared to the B1 trials. Performance measures that had little or no variability were not included for analysis.

The same organization of trials and statistical tests was applied to the over-the-shoulder ratings. B1 scores were compared to B2, B3, and B4 scores using ANOVAs and Tukey post hoc test. These are presented in Table 4. These ratings are based on an eight-point Likert scale where 1 indicates extremely poor performance and 8 indicates outstanding performance. A copy of the

**Table 4. Means and Standard Deviations for Blocks of Generic Sector Trials for Over-the-Shoulder Ratings**

VARIABLE	B1		B2		B3		B4	
	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Separating A/C	6.32	(0.92)	6.85	(0.74)	6.70	(0.68)	6.86	(0.74)
Sequencing A/C	6.07	(0.97)	6.56	(0.80)	6.60	(0.71)	6.64	(0.92)
Instructs Efficiently	6.24	(1.03)	6.41	(0.98)	6.50	(0.66)	6.83	(0.66)
Traffic Flow	6.15	(1.00)	6.53	(0.84)	6.50	(0.57)	6.69	(0.79)
Aircraft Awareness	6.06	(0.94)	6.47	(0.91)	6.33	(0.78)	6.78	(1.15)
Positive Control	6.18	(0.89)	6.51	(0.82)	6.73	(0.58)	6.81	(0.79)
Overall Awareness	6.10	(0.91)	6.56	(0.91)	6.51	(0.58)	6.89 <sup>a</sup>	(0.80)
Prioritization	6.46	(0.78)	6.56	(0.92)	6.67	(0.61)	6.92	(0.94)
Preplanning	6.21	(0.91)	6.50	(0.92)	6.50	(0.67)	6.78	(0.77)
Control Tasks	6.27	(0.90)	6.62	(0.76)	6.57	(0.67)	6.86	(0.68)
Marking Strips	6.06	(0.80)	6.56	(0.73)	6.60 <sup>a</sup>	(0.61)	6.83 <sup>a</sup>	(0.34)
Overall Prioritizing	6.12	(0.82)	6.47	(0.80)	6.50	(0.77)	6.83 <sup>a</sup>	(0.54)
Provides Ess. Info.	5.80	(0.73)	6.12	(0.89)	6.10	(0.81)	6.42 <sup>a</sup>	(0.79)
Provides Add. Info.	5.75	(0.88)	6.03	(1.04)	6.20	(0.72)	6.22	(0.84)
Overall Information	5.86	(0.74)	6.12	(0.92)	6.27	(0.82)	6.33	(0.77)
Knowledge of LOA	6.01	(1.00)	6.68 <sup>a</sup>	(0.65)	6.83 <sup>a</sup>	(0.52)	6.89 <sup>a</sup>	(0.50)
Knowledge of A/C	6.66	(0.55)	7.09 <sup>a</sup>	(0.49)	7.10 <sup>a</sup>	(0.31)	7.19 <sup>a</sup>	(0.30)
Overall Knowledge	6.44	(0.60)	6.85 <sup>a</sup>	(0.54)	6.97 <sup>a</sup>	(0.39)	7.03 <sup>a</sup>	(0.36)
Phraseology	6.86	(0.29)	6.82	(0.58)	6.90	(0.39)	6.89	(0.40)
Clear Communication	6.72	(0.48)	7.03	(0.32)	6.97	(0.27)	6.92	(0.46)
Listening to Readbacks	6.89	(0.47)	6.98	(0.58)	7.00	(0.42)	7.06	(0.24)
Overall Communication	6.77	(0.41)	6.94	(0.46)	6.90	(0.31)	6.81	(0.30)

<sup>a</sup>significantly different from Block 1 ( $p < .05$ ) as determined by the Tukey post hoc test.

questionnaire is found in Appendix B. Very few ratings were given below 6, indicating that the expertise level was already high among these controllers as a group. However, these ratings illustrate improvements in many performance rating dimensions by the time controllers executed their B4 runs. Significant improvements occurred in the controllers' ability to prioritize and provide essential information and to demonstrate a better knowledge of the LOAs for the generic sector. Items in the communication area such as phraseology, listening to readbacks, and overall communication ratings did not change.

The same organization of trials and statistical tests was applied to the controller's post-scenario questionnaire ratings and average ATWIT ratings. All comparisons were made against the B1

block of trials to assess any differences between earlier (B1) and later (B2, B3 and B4) trials. These are presented in Table 5. These ratings are based on a 1 to 10 Likert scale where 1 indicates poor performance/low workload and 10 indicates outstanding performance/high workload.

Table 5. Means and Standard Deviations for Blocks of Generic Sector Trials for Controller Self Ratings

VARIABLE	B1		B2		B3		B4	
	Mean	(SD)	Mean	(SD)	Mean	(SD)	Mean	(SD)
Overall Traffic Control	8.14	(1.07)	8.80	(1.26)	8.69	(1.11)	9.28 <sup>a</sup>	(1.02)
Workload	4.75	(1.25)	4.08	(1.55)	4.08	(1.89)	3.81	(2.31)
Communication	7.89	(1.12)	8.53 <sup>a</sup>	(0.99)	8.61 <sup>a</sup>	(1.09)	9.11 <sup>a</sup>	(0.98)
Maintaining Attention	7.44	(1.42)	8.32	(1.27)	8.19	(1.34)	8.78 <sup>a</sup>	(1.11)
Prioritization	7.75	(1.13)	8.54	(1.20)	8.36	(1.14)	8.92 <sup>a</sup>	(1.06)
Technical Knowledge	7.86	(1.23)	8.48	(1.59)	8.53	(1.52)	9.25 <sup>a</sup>	(0.94)
Safe Traffic Flow	8.11	(1.62)	9.11	(1.32)	8.81	(1.45)	9.19 <sup>a</sup>	(1.07)
Coordination	8.03	(1.13)	8.85 <sup>a</sup>	(1.02)	8.81 <sup>a</sup>	(1.09)	9.25 <sup>a</sup>	(0.83)
Improve with Practice	6.14	(2.40)	4.72 <sup>a</sup>	(3.00)	4.53 <sup>a</sup>	(3.20)	3.69 <sup>a</sup>	(3.30)
Average ATWIT	3.14	(1.39)	2.24 <sup>a</sup>	(1.05)	2.11 <sup>a</sup>	(1.16)	2.09 <sup>a</sup>	(1.20)

<sup>a</sup>significantly different from B1 ( $p < .05$ ) as determined by the Tukey post hoc test.

ATWIT ratings for each scenario were averaged across the 12 ratings made in each one-hour scenario. These ratings illustrate that the controllers perceived improvement in nearly every performance dimension and significant decrease in ATWIT ratings. There was also a significant drop in the degree to which they thought practice would improve their ability to control traffic on the generic sector.

#### 4.2.2 Orthogonal Components Analysis for Dependent Measures

Orthogonal components analyses (Buckley et al., 1983) were conducted on all dependent measures (system effectiveness variables, over-the-shoulder ratings, and controller self ratings of performance). Orthogonal components analysis looks at learning on a trial-by-trial basis and examines where performance scores change and where they begin to stabilize (Buckley et al.). This is accomplished by comparing the score on the first trial to the average of trials 2 through 6, then comparing the score on the second trial with the average of trials 3 through 6, and so on. The result is a table of probability values in which values greater than .05 are considered insignificant and values less than .05 are considered significant. Values that are .05 or less suggest that learning is still occurring. A series of values greater than .05 for a variable of interest suggests that performance has stabilized.

Stability of performance scores is important in that the variance in the scores is most likely due to the controller's true score on the variable of interest rather than the error component associated with learning. Orthogonal components analysis was conducted of the first six generic trials completed during the day 1- and day 2- training sessions.

Table 6 contains the probability values for the orthogonal components analysis of selected system effectiveness measures. The results suggest that much of the learning occurred between trial 1 and trial 2 for the majority of the system variables. With exception of the two hand-off variables (hand-offs accepted, hand-offs delayed), performance had stabilized for all of the variables by the third trial and continued to remain stable.

Table 6. Orthogonal Components Analysis for Selected System Effectiveness Variables

VARIABLE	T1 versus T2-T6	T2 versus T3-T6	T3 versus T4-T6	T4 versus T5-T6	T5 versus T6
No. of Conflicts	0.31	0.29	0.19	0.41	0.33
No. of Special Conflicts	0.0 <sup>a</sup>	0.41	0.13	0.12	0.33
No. of Sector Conflicts	0.02 <sup>a</sup>	0.49	0.13	0.08	0.08
Complexity Index	0.09	0.35	0.28	0.18	0.31
No. of Altitude Changes	0.01 <sup>a</sup>	0.34	0.48	0.19	0.22
No. of Heading Changes	0.37	0.21	0.16	0.24	0.44
No. of Speed Changes	0.04 <sup>a</sup>	0.45	0.32	0.42	0.23
PTT - Number	0.36	0.11	0.27	0.22	0.06
PTT - Duration	0.04 <sup>a</sup>	0.26	0.31	0.07	0.07
No. of Controller Key	0.05 <sup>a</sup>	0.19	0.14	0.44	0.18
No. of Sim-Pilot Key	0.47	0.20	0.25	0.19	0.24
No. of Aircraft Handled	0.11	0.19	0.49	0.06	0.29
Time Controlled (sec.)	0.02 <sup>a</sup>	0.25	0.14	0.19	0.11
Distance Flown (miles)	0.02 <sup>a</sup>	0.22	0.14	0.27	0.09
No. of Handoffs Accept	0.03 <sup>a</sup>	0.20	0.22	0.04 <sup>a</sup>	0.26
No. of Handoffs Delayed	0.03 <sup>a</sup>	0.28	0.43	0.03 <sup>a</sup>	0.45
Perc. Flights Complete	0.06	0.06	0.19	0.06	0.42
Average PTT time (sec.)	0.02 <sup>a</sup>	0.004 <sup>a</sup>	0.38	0.19	0.42
PTT/Aircraft	0.26	0.10	0.27	0.17	0.06
Conflicts/Aircraft	0.31	0.09	0.31	0.47	0.22

<sup>a</sup>significant ( $p < .05$ )

Table 7 contains the orthogonal components analysis for the over-the-shoulder ratings. The results indicate a trend similar to the system effectiveness variables. By trial three, performance began to stabilize for the majority of variables and remained stable through trial 6. This is true

Table 7. Orthogonal Components Analysis for Over-the-Shoulder Ratings

VARIABLE	T1 versus T2-T6	T2 versus T3-T6	T3 versus T4-T6	T4 versus T5-T6	T5 versus T6
Separating A/C	0.14	0.09	0.43	0.23	0.27
Sequencing A/C	0.02 <sup>a</sup>	0.28	0.43	0.48	0.21
Instructs Efficiently	0.11	0.48	0.35	0.46	0.14
Traffic Flow	0.04 <sup>a</sup>	0.36	0.48	0.46	0.12
Aircraft Awareness	0.03 <sup>a</sup>	0.42	0.33	0.39	0.27
Positive Control	0.01 <sup>a</sup>	0.45	0.24	0.31	0.35
Overall Awareness	0.002 <sup>a</sup>	0.33	0.49	0.38	0.30
Prioritization	0.10	0.27	0.45	0.25	0.08
Preplanning	0.12	0.39	0.38	0.39	0.26
Control Tasks	0.06	0.49	0.49	0.36	0.14
Marking Strips	0.04 <sup>a</sup>	0.15	0.18	0.25	0.50
Overall Prioritizing	0.05 <sup>a</sup>	0.40	0.44	0.37	0.18
Provides Ess. Info.	0.14	0.23	0.36	0.32	0.44
Provides Add. Info.	0.05 <sup>a</sup>	0.36	0.14	0.45	0.29
Overall Information	0.09	0.28	0.18	0.46	0.39
Knowledge of LOA	0.003 <sup>a</sup>	0.02 <sup>a</sup>	0.15	0.36	0.08
Knowledge of A/C	0.02 <sup>a</sup>	0.07	0.38	0.42	0.33
Overall Knowledge	0.003 <sup>a</sup>	0.05 <sup>a</sup>	0.43	0.19	0.10
Phraseology	0.22	0.15	0.24	0.47	0.15
Clear Communication	0.09	0.01 <sup>a</sup>	0.30	0.03 <sup>a</sup>	0.30
Listening to Readbacks	0.17	0.45	0.09	0.05 <sup>a</sup>	0.02 <sup>a</sup>
Overall Communication	0.25	0.10	0.20	0.14	0.01 <sup>a</sup>

<sup>a</sup>significant ( $p < .05$ )

for all variables except for several of the communication variables (clear communication, listening to readbacks, and overall communication scale rating). These variables continued to change through trials five and six indicating that learning had not stabilized completely.

Table 8 contains the orthogonal components analysis for controller self ratings including the average ATWIT score obtained for each scenario. The results are similar to the system effectiveness variables and the over-the-shoulder ratings in that, by trial 3, all measures had

Table 8. Orthogonal Components Analysis for Controller Self Ratings

VARIABLE	T1 versus T2-T6	T2 versus T3-T6	T3 versus T4-T6	T4 versus T5-T6	T5 versus T6
Overall Traffic Control	0.05 <sup>a</sup>	0.33	0.72	0.92	0.79
Workload	0.08	0.06	0.77	0.89	0.66
Communication	0.05 <sup>a</sup>	0.03 <sup>a</sup>	0.11	0.21	0.33
Maintaining Attention	0.23	0.09	0.74	0.78	0.43
Prioritization	0.03 <sup>a</sup>	0.25	0.77	0.56	0.51
Technical Knowledge	0.18	0.08	0.42	0.41	0.90
Safe Traffic Flow	0.20	0.08	0.73	0.08	0.48
Coordination	0.04 <sup>a</sup>	0.05 <sup>a</sup>	0.57	0.42	0.08
Improve with Practice	0.004 <sup>a</sup>	0.008 <sup>a</sup>	0.39	0.91	0.09
Average ATWIT	0.0005 <sup>a</sup>	0.003 <sup>a</sup>	0.02 <sup>a</sup>	0.47	0.32

<sup>a</sup>significant ( $p < .05$ )

stabilized and remained stable through trials 5 and 6. Average ATWIT ratings did not stabilize until trial 4 indicating that subjective workload was changing from trial 1 through trial 4. After trial 4, subjective workload remained at essentially the same level.

#### 4.3 Correlational Analyses for Generic and Home Sector Performance Scores

The relationship between performance scores collected on both the generic and home sectors was assessed through correlational analysis. A correlational analysis is a formal statistical technique for calculating the degree to which two variables relate or covary. The results of the analysis produce a correlation coefficient that ranges from -1.00 to +1.00 and indicates the strength and direction of the relationship between two variables. A correlation of 0.00 means no relationship exists, whereas -1.00 and +1.00 indicate a perfect relationship. A positive coefficient means that as the value of one variable increases, the value of the second variable increases as well. A negative coefficient means that as the value of one variable increases, the value of the second variable decreases. Strong positive correlation coefficients suggest that performance on the generic sector is related to performance on the home sector. Specifically, a high positive correlation indicates that if a controller performed well on a performance dimension on the generic sector, he or she also performed well on this dimension for the home sector. This same correlation would also indicate that if a controller did not perform well on a performance dimension on the generic sector, he or she also did not perform well on this dimension for the home sector.

##### 4.3.1 Reliability Analyses for Dependent Measures

The first set of correlational analyses focuses on the reliability or consistency of controller performance. A reliability analysis simply correlates one block of trials with the previous block

of trials. If the correlation is strong and significant, this indicates that performance was reliably demonstrated and measured.

One theory or assumption is that performance would be less reliable during the learning phase (i.e., earlier trials) and more reliable during the plateau or leveling off phase (i.e., later trials). For this reason, trials are arranged in a time-ordered sequence by block as in the learning and practice effect analyses.

Reliability coefficients are presented in Table 9 for selected system effectiveness variables. Significant correlation coefficients ranged from  $r = .46$  to  $r = .91$ . The results show that for most of the statistically significant variables, the magnitudes of the correlations were larger for data collected in the later trials (i.e., B3 vs. B4). This was especially noticeable in the two measures of efficiency (time under control  $r = .89$  and distance flown  $r = .89$ ).

Table 9. Reliability Analysis for Selected System Effectiveness Variables

VARIABLE	B1 VERSUS B2	B2 VERSUS B3	B3 VERSUS B4
No. of Conflicts	-0.23	0.28	-0.15
No. of Special Conflicts	0.39	-0.05	0.06
No. of Sector Conflicts	-0.10	0.43	-0.12
Complexity Index	-0.20	0.01	0.18
No. of Altitude Changes	0.46 <sup>a</sup>	0.63 <sup>a</sup>	0.57 <sup>a</sup>
No. of Heading Changes	0.15	0.01	0.20
No. of Speed Changes	0.40	0.53 <sup>a</sup>	0.07
PTT - Number	0.76 <sup>a</sup>	0.71 <sup>a</sup>	0.91 <sup>a</sup>
PTT - Duration	0.86 <sup>a</sup>	0.86 <sup>a</sup>	0.84 <sup>a</sup>
No. of Controller Key	-0.04	0.03	0.09
No. of Sim-Pilot Key	0.45 <sup>a</sup>	0.29	0.38
No. of Aircraft Handled	-0.39	-0.51 <sup>a</sup>	0.39
Time Controlled (sec.)	0.20	0.23	0.89 <sup>a</sup>
Distance Flown (miles)	0.24	0.23	0.89 <sup>a</sup>
No. of Handoffs Accept	-0.35	-0.24	0.46 <sup>a</sup>
No. of Handoffs Delayed	-0.05	-0.10	0.67 <sup>a</sup>
Perc. Flights Completed	0.48 <sup>a</sup>	0.76 <sup>a</sup>	0.73 <sup>a</sup>
Average PTT time (sec.)	0.93 <sup>a</sup>	0.91 <sup>a</sup>	0.87 <sup>a</sup>
PTT/Aircraft	0.75 <sup>a</sup>	0.73 <sup>a</sup>	0.89 <sup>a</sup>
Conflicts/Aircraft	-0.19	0.34	-0.19

<sup>a</sup>significant correlations ( $p < .05$ )

In addition, the communication variables of number and duration of PTT actions showed high reliability ( $r = .91$ ,  $r = .84$ ) in the later trials and average duration of push-to-talk ( $r = .87$ ) and number of PTT actions per aircraft ( $r = .89$ ).

Table 10 shows the results of the reliability analysis for over-the-shoulder ratings. Significant correlation coefficients ranged from  $r = .46$  to  $r = .62$ . Many of the correlations are near 0 or low in magnitude. However, for the ones that are significant, the majority occurred in comparisons involving the later trials (B2 vs. B3 or B3 vs. B4). Variables with the highest correlations were knowledge of aircraft performance capabilities, knowledge of the LOAs, and provides additional ATC information. Overall, the over-the-shoulder ratings showed low to moderate reliability.

Table 10. Reliability Analysis for Over-the-Shoulder Ratings

VARIABLE	B1 VERSUS B2	B2 VERSUS B3	B3 VERSUS B4
Separating A/C	0.22	-0.36	-0.24
Sequencing A/C	0.03	-0.31	-0.06
Instructs Efficiently	0.23	0.07	0.46 <sup>a</sup>
Traffic Flow	0.21	-0.13	0.11
Aircraft Awareness	0.31	-0.19	0.23
Positive Control	0.47 <sup>a</sup>	-0.07	0.00
Overall Awareness	0.33	-0.02	-0.11
Prioritization	0.17	-0.11	0.10
Preplanning	0.44	-0.13	0.16
Control Tasks	0.44	-0.10	0.10
Marking Strips	0.36	0.54 <sup>a</sup>	-0.11
Overall Prioritizing	0.17	-0.07	0.06
Provides Ess. Info.	0.50 <sup>a</sup>	0.34	0.39
Provides Add. Info.	0.33	0.46 <sup>a</sup>	0.62 <sup>a</sup>
Overall Information	0.41	0.46 <sup>a</sup>	0.36
Knowledge of LOA	0.50 <sup>a</sup>	0.58 <sup>a</sup>	0.47 <sup>a</sup>
Knowledge of A/C	0.21	0.44	0.60 <sup>a</sup>
Overall Knowledge	0.59 <sup>a</sup>	0.43	0.40
Phraseology	0.49 <sup>a</sup>	0.21	-0.07
Clear Communication	0.27	0.18	-0.01
Listening to Readbacks	0.06	-0.06	-0.06
Overall Communication	0.45 <sup>a</sup>	0.09	-0.26

<sup>a</sup>statistically significant ( $p < .05$ )



Table 11 shows the reliability coefficients for the controller self ratings of performance. Significant correlation coefficients ranged from  $r = .48$  to  $r = .97$ . This data set shows a marked increase in the size of the correlations in the later trials compared to the earlier trials with correlations for all variables significant in the B3 vs. B4 comparisons. Variables with the highest correlations include communication ( $r = .94$ ), coordination with others, ( $r = .88$ ) and average ATWIT ratings ( $r = .97$ ). Overall, the controller self ratings of performance showed high reliability.

Table 11. Reliability Analysis For Controller Self Ratings

VARIABLE	B1 VERSUS B2	B2 VERSUS B3	B3 VERSUS B4
Overall Traffic Control	0.38	0.38	0.60 <sup>a</sup>
Workload	0.56 <sup>a</sup>	0.57 <sup>a</sup>	0.77 <sup>a</sup>
Communication	0.69 <sup>a</sup>	0.74 <sup>a</sup>	0.94 <sup>a</sup>
Maintaining Attention	0.14	0.60 <sup>a</sup>	0.56 <sup>a</sup>
Prioritization	0.23	0.28	0.66 <sup>a</sup>
Technical Knowledge	0.55 <sup>a</sup>	0.63 <sup>a</sup>	0.62 <sup>a</sup>
Safe Traffic Flow	0.29	0.31	0.48 <sup>a</sup>
Coordination	0.36	0.54 <sup>a</sup>	0.88 <sup>a</sup>
Improve with Practice	0.83 <sup>a</sup>	0.92 <sup>a</sup>	0.92 <sup>a</sup>
Average ATWIT	0.81 <sup>a</sup>	0.94 <sup>a</sup>	0.97 <sup>a</sup>

<sup>a</sup>statistically significant ( $p < .05$ )

#### 4.3.2 Correlational Relationships Between Generic and Home Sector Performance Scores

The relationship between performance on the generic sector and performance on the home sector was assessed through correlational analysis. Scores collected from the day-1, -2, and -3 generic traffic runs were correlated with scores collected from the day-3 home sector traffic runs. This was done for all performance categories (system variables, over-the-shoulder ratings, controller self ratings, and ATWIT). High correlations between generic and home-sector scores would indicate that controllers, as a group, tend to perform in a similar fashion on the generic sector as they would on their home sector for that dimension. Low correlations could indicate a number of things including the possibility that measurement was not reliable enough to allow the presence of a significant correlational relationship between home and generic sectors. The data for the following analysis are arranged by block as in the learning analysis. The correlations in each block represent the relationship between the average of the day-3 home sector runs and the average of the generic runs for that particular block of trials.

Table 12 shows the correlations between home and generic sectors for selected system effectiveness measures. Significant correlation coefficients ranged from  $r = .45$  to  $r = .87$ . The general trend of more significant and higher correlations in the later runs is evident in this data set. High and significant correlations were found for measures of communication activity

Table 12. Correlation Between Home Sector and Generic Sector Blocks for Selected System Effectiveness Variables

VARIABLE	HOME VERSUS B1	HOME VERSUS B2	HOME VERSUS B3	HOME VERSUS B4
No. of Conflicts	0.23	-0.23	-0.05	0.46 <sup>a</sup>
No. of Special Conflicts	-0.21	0.04	0.12	0.23
No. of Sector Conflicts	-0.10	-0.14	-0.13	-0.12
Complexity Index	0.65	0.04	0.30	-0.02
No. of Altitude Changes	0.28	-0.05	0.22	0.22
No. of Heading Changes	0.39	0.30	0.23	0.61 <sup>a</sup>
No. of Speed Changes	0.44	0.52 <sup>a</sup>	0.47 <sup>a</sup>	0.44
PTT - Number	0.63 <sup>a</sup>	0.77 <sup>a</sup>	0.60 <sup>a</sup>	0.62 <sup>a</sup>
PTT - Duration	0.82 <sup>a</sup>	0.87 <sup>a</sup>	0.81 <sup>a</sup>	0.73 <sup>a</sup>
No. of Controller Key	0.13	0.09	0.19	0.03
No. of Sim-Pilot Key	0.23	0.55 <sup>a</sup>	0.33	0.69 <sup>a</sup>
No. of Aircraft Handled	0.31	-0.05	0.23	0.12
Time Controlled (sec.)	0.45 <sup>a</sup>	0.21	0.82 <sup>a</sup>	0.74 <sup>a</sup>
Distance Flown (miles)	0.43	0.22	0.80 <sup>a</sup>	0.76 <sup>a</sup>
No. of Handoffs Accept	0.41	-0.22	0.27	0.25
No. of Handoffs Delayed	0.16	-0.21	0.15	0.25
Perc. Flights Completed	-0.14	0.18	0.17	0.09
Average PTT time (sec.)	0.78 <sup>a</sup>	0.86 <sup>a</sup>	0.85 <sup>a</sup>	0.76 <sup>a</sup>
PTT/Aircraft	0.58 <sup>a</sup>	0.78 <sup>a</sup>	0.55 <sup>a</sup>	0.61 <sup>a</sup>
Conflicts/Aircraft	0.25	-0.23	-0.06	0.29

<sup>a</sup>statistically significant ( $p < .05$ ), (degrees of freedom = 17)

(number and duration of PTT) and measures of efficiency (time under control and distance flown). Significant correlations were present for number of conflicts in the B4 runs. Overall, these variables showed that, although performance in generic airspace was not a perfect analog to that in the home sector, it did provide many similarities.

Table 13 shows the correlations between home and generic sectors for over-the-shoulder ratings of performance. Correlation coefficients ranged from  $r = .46$  to  $r = .63$ . Unlike the previous analysis of system variables, the over-the-shoulder ratings showed more significant correlations during the B2 runs compared to the B4 runs. Many significant correlations were present in the providing information variable dimensions (providing essential information, providing additional

Table 13. Correlation Between Home Sector and Generic Sector Blocks for Over-the-Shoulder Ratings

VARIABLE	HOME VERSUS B1	HOME VERSUS B2	HOME VERSUS B3	HOME VERSUS B4
Separating A/C	0.52 <sup>a</sup>	0.34	-0.39	0.12
Sequencing A/C	0.23	0.22	-0.35	0.19
Instructs Efficiently	0.40	0.19	-0.16	-0.26
Traffic Flow	0.50 <sup>a</sup>	0.43	-0.13	0.26
Aircraft Awareness	0.40	0.44	0.04	-0.02
Positive Control	0.40	0.24	-0.20	0.07
Overall Awareness	0.19	0.19	-0.11	-0.04
Prioritization	0.26	0.44	-0.24	-0.06
Preplanning	0.36	0.49 <sup>a</sup>	-0.26	0.22
Control Tasks	0.15	-0.01	-0.26	0.13
Marking Strips	0.22	0.52 <sup>a</sup>	0.07	-0.02
Overall Prioritizing	0.31	0.56 <sup>a</sup>	-0.43	0.08
Provides Ess. Info.	0.23	0.62 <sup>a</sup>	0.33	0.33
Provides Add. Info.	0.42	0.63 <sup>a</sup>	0.62 <sup>a</sup>	0.70 <sup>a</sup>
Overall Information	0.31	0.62 <sup>a</sup>	0.39	0.58 <sup>a</sup>
Knowledge of LOA	0.22	0.17	0.32	-0.20
Knowledge of A/C	0.71 <sup>a</sup>	0.35	0.28	0.46 <sup>a</sup>
Overall Knowledge	0.39	0.49 <sup>a</sup>	0.47 <sup>a</sup>	0.25
Phraseology	0.63 <sup>a</sup>	0.44	0.32	0.47 <sup>a</sup>
Clear Communication	0.24	0.24	0.04	-0.04
Listening to Readbacks	0.21	0.14	0.24	0.00
Overall Communication	0.58 <sup>a</sup>	0.50 <sup>a</sup>	0.03	0.58

<sup>a</sup>statistically significant ( $p < .05$ ), (degrees of freedom = 17)

information, overall information) for the B2 runs. In addition, the prioritization variables (marking flight strips, preplanning, and overall prioritization) showed significant correlations in the B2 runs. Overall, the over-the-shoulder ratings showed low correlations between home and generic sector performance. However, this could be, in part, a function of the reliability of the rating as indicated in Table 10. Performance observation and rating are inherently difficult. The rating form used in this study was an earlier version. There is an ongoing program to improve it along with the rater training package that must accompany it.

Table 14 shows the correlations between home and generic sectors for controller self ratings of performance and ATWIT. Correlation coefficients ranged from  $r = .48$  to  $r = .95$ . By B4, all variable dimensions showed high and significant correlations between generic and home sector runs. High correlations were obtained for improvement with practice ( $r = .94$ ), technical knowledge ( $r = .76$ ) and prioritization ( $r = .76$ ). Average ATWIT ratings also showed very high correlations for all blocks of trials. Average ATWIT correlation coefficients ranged from  $r = .66$  for the B1 runs to  $r = .95$  for the B4 runs. These results suggest that workload was very similar for working home versus generic sector scenarios for controllers as a group. It is also noteworthy that ATWIT ratings produced the highest reliabilities and the highest between-sector correlations of all the controller self-rating variables.

Table 14. Correlation Between Home Sector and Generic Sector Blocks for Controller Self Ratings

VARIABLE	HOME VERSUS B1	HOME VERSUS B2	HOME VERSUS B3	HOME VERSUS B4
Overall Traffic Control	0.41	0.48 <sup>a</sup>	0.64 <sup>a</sup>	0.69 <sup>a</sup>
Workload	0.56 <sup>a</sup>	0.72 <sup>a</sup>	0.85 <sup>a</sup>	0.75 <sup>a</sup>
Communication	0.60 <sup>a</sup>	0.58 <sup>a</sup>	0.74 <sup>a</sup>	0.65 <sup>a</sup>
Maintaining Attention	0.20	0.39	0.54 <sup>a</sup>	0.63 <sup>a</sup>
Prioritization	0.37	0.31	0.74 <sup>a</sup>	0.76 <sup>a</sup>
Technical Knowledge	0.55 <sup>a</sup>	0.66 <sup>a</sup>	0.83 <sup>a</sup>	0.76 <sup>a</sup>
Safe Traffic Flow	0.53 <sup>a</sup>	0.49 <sup>a</sup>	0.76 <sup>a</sup>	0.63 <sup>a</sup>
Coordination	0.54 <sup>a</sup>	0.60 <sup>a</sup>	0.61 <sup>a</sup>	0.49 <sup>a</sup>
Improve with Practice	0.83 <sup>a</sup>	0.87 <sup>a</sup>	0.97 <sup>a</sup>	0.94 <sup>a</sup>
Average ATWIT	0.66 <sup>a</sup>	0.88 <sup>a</sup>	0.93 <sup>a</sup>	0.95 <sup>a</sup>

<sup>a</sup> statistically significant ( $p < .05$ ), (degrees of freedom = 17)

#### 4.3.3 Correlational Analyses Between ATWIT Ratings and Controller Self Ratings of Performance

This section deals with the relationship between controller workload and controller performance. The primary measure of workload in this study was ATWIT ratings taken at 5-minute intervals and then averaged producing a score for each scenario.

The most compatible performance data set was the controller self ratings of performance taken immediately after each scenario was completed. Table 15 presents the correlational relationships between average ATWIT ratings and controller self ratings of performance. The scores are arranged by block as in the earlier analyses with the inclusion of a Block 5 (B5) of scenarios. B5 represents the average of the two home-sector runs performed on day 3 of testing.

Table 15. Correlation Between Average ATWIT Scores and Controller Self Ratings of Performance

VARIABLE	B1	B2	B3	B4	B5 <sup>a</sup>
Overall Traffic Control	-0.20	-0.62 <sup>b</sup>	-0.58 <sup>b</sup>	-0.70 <sup>b</sup>	-0.74 <sup>a</sup>
Workload	0.25	0.66 <sup>b</sup>	0.89 <sup>b</sup>	0.78 <sup>b</sup>	0.84 <sup>a</sup>
Communication	-0.22	-0.74 <sup>b</sup>	-0.80 <sup>b</sup>	-0.72 <sup>b</sup>	-0.64 <sup>a</sup>
Maintaining Attention	-0.26	-0.25	-0.47 <sup>b</sup>	-0.44	-0.35
Prioritization	-0.36	-0.33	-0.69 <sup>b</sup>	-0.65 <sup>b</sup>	-0.76 <sup>a</sup>
Technical Knowledge	-0.06	-0.20	-0.47 <sup>b</sup>	-0.59 <sup>b</sup>	-0.50 <sup>a</sup>
Safe Traffic Flow	-0.50 <sup>b</sup>	-0.58 <sup>b</sup>	-0.63 <sup>b</sup>	-0.72 <sup>b</sup>	-0.86 <sup>a</sup>
Coordination	-0.53 <sup>b</sup>	-0.37	-0.76 <sup>b</sup>	-0.65 <sup>b</sup>	-0.39
Improve with Practice	0.57 <sup>b</sup>	0.78 <sup>b</sup>	0.84 <sup>b</sup>	0.77 <sup>b</sup>	0.83 <sup>a</sup>

<sup>a</sup>average of Home 3 and Home 4

<sup>b</sup>statistically significant ( $p < .05$ )

The results show negative correlations between ATWIT ratings and controller self ratings of performance. This trend holds regardless of whether the ATWIT ratings were taken from a home or generic sector scenario as illustrated by the similarity between the B5 (home) correlations and the B4 (generic) correlation coefficients. All the correlations were negative with exception of the workload ratings and improvement with practice. The explanation for the negative correlations is that controllers with the best performance (as measured by their scale rating) gave the lowest ATWIT ratings. One interpretation for the positive correlation for the practice improvement scale rating is that controllers with the least need for practice also gave the lowest ATWIT ratings. The positive correlation for the workload rating indicates an agreement between the controller's average ATWIT rating and his or her overall workload rating made at the end of the scenario. These results are in line with previous results on workload assessment such as in Stein (1985).

#### 4.4 Final Questionnaire Comments on the Entire Experiment

A final questionnaire was administered to each controller at the end of their day-3 session. The questions requested information concerning the realism of the simulation, the representativeness of the generic sector, and the effectiveness of the training aids used. Table 16 summarizes these comments. As far as the realism of the simulation, the majority (13 of 17) of the controllers thought the simulation was moderately realistic or better. Three controllers thought

Table 16. Summary of Controller Final Questionnaire Comments

Realism of the simulation	Controllers Commenting
Very realistic	3
Moderately realistic	10
Somewhat realistic	3
Not realistic	1
Representativeness of a typical sector	Controllers Commenting
Very representative	8
Close to typical	6
Somewhat representative	2
Not typical	1
Helpfulness of the training booklet	Controllers Commenting
Very helpful	4
Helpful	4
Somewhat helpful	4
Not helpful	3
Didn't read it	2
Adequacy of the hands-on-training	Controllers Commenting
Yes, it was adequate	17
No, it was not adequate	0
Responsiveness of the simulation pilots	Controllers Commenting
Excellent job	3
Very well	7
Good job	6
Not very good	1
Intrusiveness of the ATWIT device	Controllers Commenting
No, not at all	16
Yes, it was intrusive	0

the simulation was somewhat realistic and one controller did not think the simulation was realistic. The majority of the controllers thought that the generic sector was representative of a typical sector (14 of 17). Two controllers thought the sector was somewhat representative and one controller thought it was not typical. Most controllers stated that the training manual was helpful. They thought the map and frequencies for adjacent sectors were the most useful pieces of information in the booklet. All controllers responded positively to the hands-on training they received during the day-1 and day-2 sessions. Appendix C lists the questions and a complete transcript of responses.

## 5. Summary and Conclusions

Eight system effectiveness variables were significantly lower by the B4 runs compared to the B1 runs. Many of these measures coincide with three of four factors derived from Buckley's two experiments (Buckley et al., 1983). These factors and significant measures from this experiment include confliction (special conflicts and traffic complexity), occupancy (time under control, distance flown, and percentage of flights completed), and communication (number of heading and speed changes and average PTT time).

Over-the-shoulder ratings for the learning trials indicated that, by the fourth block of trials, controllers performed better on a number of rating variables. This also supports the findings with

the ATWIT ratings that after controllers learned the fixes, airways, and typical flight plans, they were better able to perform control tasks such as flight strip marking and providing information to the simulation pilots. Ratings were also higher for SA variables such as awareness of aircraft positions. Lastly, there were positive indications that controllers had learned and applied the LOAs effectively by the last block of generic runs. Orthogonal components analysis showed that 19 of 22 of these variables had stabilized by the fourth generic run.

Controller self ratings showed a similar trend as the system variables and over-the-shoulder ratings. By the fourth block of runs, nearly every rating variable showed a performance improvement compared to the first block of runs. Controllers felt that their overall ability to control traffic was significantly better by the later runs especially in the areas of applying technical knowledge, maintaining a safe and efficient traffic flow, and coordinating with others. They also felt there was little need for additional practice by the last block of runs.

### 5.1 Discussion of Learning Rate for the Generic Sector

Learning rate for the generic sector can be inferred from differences in the performance scores over trials on the first 2 days of testing. These scores were collected from the four performance measurement categories (system effectiveness variables, over-the-shoulder ratings, post-scenario questionnaire ratings, and ATWIT ratings). ATWIT ratings provided the strongest support for learning with significantly lower scores by the last trial. Orthogonal components analysis provided a more detailed view of the learning curve and showed that by the fourth trial, ATWIT ratings had begun to level off and plateau. One explanation for these findings is that many features of the sector became more familiar as controllers went through the multiple generic runs. Specifically, controllers learned the fix locations, the airways, the typical flight plans, and crossing restrictions. As this information was learned, it became more automatic, and the controller did not have to expend as much energy thinking about these sector features as they did during the initial runs.

### 5.2 Discussion of Correlational Relationships Between Performance Scores

#### 5.2.1 Discussion of Reliability of Performance Scores

Reliability of performance scores varied quite a bit among the four categories of performance scores. However, all categories showed improvements in reliability towards the later trials. ATWIT ratings were the most reliable of all the measures demonstrating almost perfect reliability. Controller self ratings were next, showing reliability across trials. The system variables were next with somewhat inconsistent reliability, and the over-the-shoulder ratings were last with fairly low reliability.

Variations in reliability can arise from a number of reasons. First, performance can actually fluctuate causing variations from trial to trial. Second, the measurement of performance can fluctuate causing variations from trial to trial. The question remains as to why ATWIT ratings show almost perfect reliability, whereas over-the-shoulder ratings show poor reliability. Differences in measurement of ATWIT versus over-the-shoulder ratings almost certainly caused this difference. ATWIT scores are based on controller self ratings at 5-minute intervals during the scenario. Each scenario is one hour in length, therefore, the average ATWIT rating is based



on 12 recorded observations. Over-the-shoulder ratings were made once at the end of the scenario and were based on an unknown number of observations. ATWIT ratings were made more frequently, therefore, the reliability of the ratings was higher. Also, measurement reliability is far from perfect and undoubtedly varies across measurement tools.

#### 5.2.2 Discussion of Correlations Between Performance Scores on the Home Sector and Generic Sector

Three of the four performance categories showed high and consistent correlations between the generic and home sectors. These categories were ATWIT ratings, system effectiveness measures, and controller self ratings of performance. These correlations suggest that controller workload, communication, and task management were basically similar regardless of the sector configuration. Workload, as measured by ATWIT, was also highly correlated between the home sector and the fourth block of generic runs. This result suggests that once the sector was learned, the workload was the same regardless of the sector configuration. The results also indicate that system performance, as measured by system effectiveness measures, was very similar in both sector configurations.

The over-the-shoulder ratings showed low correlations between home and generic sectors. This could mean that there are low relationships between rating dimensions for the two sectors. Given the fact that the majority of the other data does correlate, a more likely hypothesis is that there are some measurement issues associated with the collection of over-the-shoulder rating data. The fact that only one observer-rater was available probably complicated this issue. This hypothesis is further supported by the low reliability found for the over-the-shoulder ratings. True correlations may exist, but the low reliability of measurement may be obscuring these relationships. One possible solution to increase the reliability of these ratings is to have the rater make ratings at intervals during the scenario. In this method, ratings could take place at perhaps 10-minute intervals. The rating would only be based on observations that occurred during that interval. A single score for the scenario could be calculated by deriving an average performance measurement score for each rating dimension used.

It is likely that all of the views expressed about human performance have some merit in their own right. We need to look at how human beings behave in complex systems from a variety of perspectives. These include those that focus on basic psychological functions and those that center on very task-specific issues. The latter concept can include assessment of molecular variables in an automated and objective laboratory environment and SME ratings, if done in a systematic and objective fashion.

The research approaches in the FAA RDHFL leave the issue open. The ultimate goal is to learn how people perform under often demanding task load so that we can ultimately help them do it better with a decreased probability of human error.

This has been the second in a series of studies examining the efficacy of using generic airspace in real time simulation. This study has been consistent with past findings indicating that generic airspace is a viable tool for system test and evaluation. The information in this test supports the notion that en route controllers can quickly learn a generic airspace and that performance in a generic airspace is related to performance on a home sector. The use of generic airspace will



allow human factors researchers to more easily generalize to the population of air traffic controllers by conducting tests on a standardized airspace.

## References

- Bailey, R. W. (1982). *Human performance engineering: A guide for system designers*. Englewood Cliffs, NJ: Prentice Hall.
- Berlinger, D. C., Angell, D., & Shearer, J. W. (1964). Behavior, measures instruments for performance evaluation in simulated environments. *Proceedings of the Symposium and Workshop on the Quantification of Human Performance*, pp. 227-296.
- Buckley, E. P., DeBaryshe, B. D., Hitchner, N., & Kohn, P. (1983). *Methods and measurements in real-time air traffic control system simulation* (DOT/FAA/CT-83/26). Atlantic City, NJ: DOT/FAA Technical Center.
- Buckley, E. P., O'Connor, W. F., Beebe, T., Adams, W., & MacDonald, G. (1969). A comparative analysis of individual and system performance indices for the air traffic control system (Report No. FAA-NA-69-40). Atlantic City, NJ: National Aviation Facilities Experimental Center. (NTIS No. AD-710 795)
- Endsley, M. R., & Kiris, E. O. (1995). The out of the loop performance problem and level of control in automation, *Human Factors*, 37(2), 381-394.
- Endsley, M. R., & Rodgers, M. D. (1994). Situation awareness information requirements analysis for en route air traffic control. *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting* (pp. 73-75). Santa Monica, CA: Human Factors Society.
- FAA (1988). *Profile of operational errors in the national airspace system calendar year 1987*, Washington DC: Office of Aviation Safety, Safety Information, and Analysis Division.
- FAA (1990). *ATCT/ARTCC OJT instruction/evaluation report* (FAA Form 3120-25). Washington, DC: Federal Aviation Administration.
- FAA (1996). *Administrator's fact book*. Federal Aviation Administration (ABC-100): Washington, DC.
- Flach, J. W. (1995). Situation awareness: proceed with caution. *Human Factors*, 37(1), 149-157.
- Gay, L. R. (1994). *Educational research: Competencies for analysis and application* (4<sup>th</sup> edition). Columbus, OH: Merrill.
- Guttman, J. A., Stein, E. S., & Gromelski, S. (1995). *The influence of generic airspace on air traffic controller performance* (DOT/FAA/CT-TN95/38). Atlantic City, NJ: DOT/FAA Technical Center.
- Hopkin, V. D. (1991). The impact of automation on air traffic control systems. In J. A. Wise. *Automation and systems issues in air traffic control* (pp. 3-19). Berlin: Springer-Verlag.

- Kinney, G. C., Spahn, M. J., & Amato, R. A. (1977). *The human element in air traffic control: Observations and analysis of the performance of controllers and supervisors in providing ATC separation services* (Report No. MTR-7655). McLean, VA: The MITRE Corporation.
- Mogford, R. H., Murphy, E. D., Yastrop, G., Guttman, J. A. & Roske-Hofstrand, R. J.(1993). *The application of research techniques for documenting cognitive processes in air traffic control* (Report No. DOT/FAA/CT-TN93/39). Atlantic City, NJ: Federal Aviation Administration.
- Paul, L. E. (1989). *The evaluation of conflicts in air traffic control simulation*. Unpublished manuscript. DOT/FAA Technical Center, Atlantic City, NJ.
- Paul, L. E. (1990). *Using simulation to evaluate the safety of proposed ATC operations and procedures* (DOT/FAA/CT-TN90/22). Atlantic City, NJ: Federal Aviation Administration Technical Center.
- Rodgers, M. D. (1993). *An examination of the operational error data base for air traffic control centers* (DOT/FAA/AM-93/12). Oklahoma City: FAA Civil Aeromedical Institute.
- Rodgers, M. D. & Duke, D. A. (1993). *SATORI: Situation assessment through the recreation of incidents* (DOT/FAA/AM-93/12). Oklahoma City: FAA Civil Aeromedical Institute.
- Rodgers, M. D., Manning C. A., & Kerr, C. S. (1994). Demonstration of power: Performance and objective workload evaluation research. *Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting* (p. 941). San Diego, CA: Human Factors and Ergonomics Society.
- Seamster, T. L., Redding, R. E., Canon, J. R., Ryder, J. M., & Purcell, J. A. (1993). Cognitive task analysis of expertise in air traffic control. *The International Journal of Aviation Psychology*, 3(4), 257-283.
- Sollenberger, R. L., & Stein, E. S. (1995). *The effects of structured arrival and departure procedures on TRACON air traffic controller memory and situational awareness* (DOT/FAA/CT-TN95/27). Atlantic City, NJ: DOT/FAA Technical Center.
- Sollenberger, R. L., Stein, E. S., & Gromelski, S. (1997). *The development and evaluation of a behaviorally based rating form for the assessment of air traffic controller performance* (DOT/FAA/CT-TN96/16). Atlantic City, NJ: DOT/FAA Technical Center.
- Stein, E. S. (1984a). *The measurement of pilot performance - a master-journeyman approach* (Report No. DOT/FAA/CT-83/15). Atlantic City, NJ: DOT/FAA Technical Center.
- Stein, E. S. (1984b). Observing rating of air traffic controller workload during simulation. In V. Amico and A. B. Clymer (Eds.), *Proceedings of the 1984 SCS Simulators Conference*, 14(1), 288-290.
- Stein, E. S. (1985). *Air traffic controller workload: An examination of workload probe* (Report No. DOT/FAA/CT-TN 84/24). Atlantic City, NJ: DOT/FAA Technical Center.

Stein, E. S. (1989). *Parallel approach separation and controller performance* (Report No. DOT/FAA/CT-TN 89/50). Atlantic City, NJ: DOT/FAA Technical Center.

Stein E. S. & Buckley, E. P. (1992). *Simulation variables*. Unpublished manuscript.

Thorndike, R. L.. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

Zingale, C., Gromelski, S., Ahmed, B., & Stein, E. S. (1993). *Influence of individual experience and flight strips on air traffic controller memory/situational awareness* (Report No. DOT/FAA/CT-TN93/21). Atlantic City, NJ DOT/FAA Technical Center.

Appendix A  
Demographic Form

I.D. # \_\_\_\_\_

1) What is your age in years?

\_\_\_\_\_ years

2) How many years have you actively controlled traffic?

\_\_\_\_\_ years

3) How many years have you controlled traffic at the Jacksonville ARTCC?

\_\_\_\_\_ years

4) How many months in the past year have actively controlled traffic?

\_\_\_\_\_ months

5) What is your current position as an air traffic controller?

☐ Developmental      ☐ Full Performance Level   ☐ Other \_\_\_\_\_

6) Are you wearing corrective lenses during this test?

☐ Yes   ☐ No

7) Circle the number which best describes your current skill as an air traffic controller.

1   2   3   4   5   6   7   8   9   10

Not very skilled

Extremely skilled

8) Circle the number which best describes your motivation to participate in this study.

1   2   3   4   5   6   7   8   9   10

Not very motivated

Extremely motivated

9) Circle the number which best describes your current state of health

1   2   3   4   5   6   7   8   9   10

Not very healthy

Extremely healthy

10) Please indicate the frequency that you play video games. \_\_\_\_\_ hours per month

Appendix B  
Observer Evaluation Form

Date \_\_\_\_\_

Controller \_\_\_\_\_

Sector            JAX                      GEN

**INSTRUCTIONS**

This form was designed to be used by instructor certified Air Traffic Control Specialists to evaluate the effectiveness of controllers working in simulation environments. Observers will rate the effectiveness of controllers in several different performance areas using the scale shown below. When making your ratings, please try to use the entire scale range as much as possible. You are encouraged to write down observations and you may make preliminary ratings during the course of the scenario. However, we recommend that you wait until the scenario is finished before making your final ratings. The observations you make do not need to be restricted to the performance areas covered in this form and may include other areas that you think are important. Also, please write down any comments that may improve this evaluation form. Your identity will remain anonymous, so do not write your name on the form. Instead, your data will be identified by an observer code known only to yourself and the researchers conducting this study.

Rating	Scale Point Description
1	Controller demonstrated <i>extremely</i> poor judgment in making control decisions and <i>very</i> frequently made errors
2	Controller demonstrated poor judgment in making some control decisions and occasionally made errors
3	Controller made questionable control decisions using poor control techniques which led to restricting the normal traffic flow
4	Controller demonstrated the ability to keep aircraft separated but used spacing and separation criteria which was excessive
5	Controller demonstrated <i>adequate</i> judgment in making control decisions
6	Controller demonstrated <i>good</i> judgment in making control decisions using efficient control techniques
7	Controller <i>frequently</i> demonstrated <i>excellent</i> judgment in making control decisions using extremely good control techniques
8	Controller <i>always</i> demonstrated excellent judgment in making even the most difficult control decisions while using outstanding control techniques
NA	Not Applicable - There was not an opportunity to observe performance in this particular area during the simulation

## I - MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW

1. Maintaining Separation and Resolving Potential Conflicts..... 1 2 3 4 5 6 7 8 NA
- using control instructions that maintain safe aircraft separation
  - detecting and resolving impending conflicts early
2. Sequencing Arrival and Departure Aircraft Efficiently..... 1 2 3 4 5 6 7 8 NA
- using efficient and orderly spacing techniques for arrival and departure aircraft
  - maintaining safe arrival and departure intervals that minimize delays
3. Using Control Instructions Effectively..... 1 2 3 4 5 6 7 8 NA
- providing accurate navigational assistance to pilots
  - avoiding clearances that result in the need for additional instructions to handle aircraft completely
  - avoiding excessive vectoring or over-controlling
4. Overall Safe and Efficient Traffic Flow Scale Rating..... 1 2 3 4 5 6 7 8 NA

## II - MAINTAINING ATTENTION AND SITUATION AWARENESS

5. Maintaining Awareness of Aircraft Positions ..... 1 2 3 4 5 6 7 8 NA
- avoiding fixation on one area of the radar scope when other areas need attention
  - using scanning patterns that monitor all aircraft on the radar scope
6. Ensuring Positive Control..... 1 2 3 4 5 6 7 8 NA
7. Detecting Pilot Deviations from Control Instructions..... 1 2 3 4 5 6 7 8 NA
- ensuring that pilots follow assigned clearances correctly
  - correcting pilot deviations in a timely manner
8. Correcting Own Errors in a Timely Manner ..... 1 2 3 4 5 6 7 8 NA
9. Overall Attention and Situation Awareness Scale Rating ..... 1 2 3 4 5 6 7 8 NA

### III - PRIORITIZING

10. Taking Actions in an Appropriate Order of Importance ..... 1 2 3 4 5 6 7 8 NA
- resolving situations that need immediate attention before handling low priority tasks
  - issuing control instructions in a prioritized, structured, and timely manner
11. Preplanning Control Actions..... 1 2 3 4 5 6 7 8 NA
- scanning adjacent sectors to plan for inbound traffic
  - studying pending flight strips in bay
12. Handling Control Tasks for Several Aircraft ..... 1 2 3 4 5 6 7 8 NA
- shifting control tasks between several aircraft when necessary
  - avoiding delays in communications while thinking or planning control actions
13. Marking Flight Strips while Performing Other Tasks ..... 1 2 3 4 5 6 7 8 NA
- marking flight strips accurately while talking or performing other tasks
  - keeping flight strips current
14. Overall Prioritizing Scale Rating ..... 1 2 3 4 5 6 7 8 NA

### IV - PROVIDING CONTROL INFORMATION

15. Providing Essential Air Traffic Control Information ..... 1 2 3 4 5 6 7 8 NA
- providing mandatory services and advisories to pilots in a timely manner
  - exchanging essential information
16. Providing Additional Air Traffic Control Information ..... 1 2 3 4 5 6 7 8 NA
- providing additional services when workload is not a factor
  - exchanging additional information
17. Overall Providing Control Information Scale Rating..... 1 2 3 4 5 6 7 8 NA



## V - TECHNICAL KNOWLEDGE

18. Showing Knowledge of LOAs and SOPs..... 1 2 3 4 5 6 7 8 NA
- controlling traffic as depicted in current LOAs and SOPs
  - performing handoff procedures correctly
19. Showing Knowledge of Aircraft Capabilities and Limitations..... 1 2 3 4 5 6 7 8 NA
- avoiding clearances that are beyond aircraft performance parameters
  - recognizing the need for speed restrictions and wake turbulence separation
20. Overall Technical Knowledge Scale Rating ..... 1 2 3 4 5 6 7 8 NA

## VI - COMMUNICATING

21. Using Proper Phraseology..... 1 2 3 4 5 6 7 8 NA
- using words and phrases specified in ATP 7110.65
  - using ATP phraseology that is appropriate for the situation
  - avoiding the use of excessive verbiage
22. Communicating Clearly and Efficiently..... 1 2 3 4 5 6 7 8 NA
- speaking at the proper volume and rate for pilots to understand
  - speaking fluently while scanning or performing other tasks
  - clearance delivery is complete, correct and timely
  - providing complete information in each clearance
23. Listening to Pilot Readbacks and Requests..... 1 2 3 4 5 6 7 8 NA
- correcting pilot readback errors
  - acknowledging pilot or other controller requests promptly
  - processing requests correctly in a timely manner
24. Overall Communicating Scale Rating ..... 1 2 3 4 5 6 7 8 NA

## **I - MAINTAINING SAFE AND EFFICIENT TRAFFIC FLOW**

### **1. Maintaining Separation and Resolving Potential Conflicts**

### **2. Sequencing Arrival and Departure Aircraft Efficiently**

### **3. Using Control Instructions Effectively**

### **4. Other Actions Observed in Safe and Efficient Traffic Flow**

## **II - MAINTAINING ATTENTION AND SITUATION AWARENESS**

### **5. Maintaining Awareness of Aircraft Positions**

### **6. Ensuring Positive Control**

### **7. Detecting Pilot Deviations from Control Instructions**

### **8. Correcting Own Errors in a Timely Manner**

### **9. Other Actions Observed in Attention and Situation Awareness**

### III - PRIORITIZING

10. Taking Actions in an Appropriate Order of Importance

11. Preplanning Control Actions

12. Handling Control Tasks for Several Aircraft

13. Marking Flight Strips while Performing Other Tasks

14. Other Actions Observed in Prioritizing

### IV - PROVIDING CONTROL INFORMATION

15. Providing Essential Air Traffic Control Information

16. Providing Additional Air Traffic Control Information

17. Other Actions Observed in Providing Control Information

## **V - TECHNICAL KNOWLEDGE**

**18. Showing Knowledge of LOAs and SOPs**

**19. Showing Knowledge of Aircraft Capabilities and Limitations**

**20. Other Actions Observed in Technical Knowledge**

## **VI - COMMUNICATING**

**21. Using Proper Phraseology**

**22. Communicating Clearly and Efficiently**

**23. Listening to Pilot Readbacks and Requests**

**24. Other Actions Observed in Communicating**

Appendix C  
Transcript of Controller Final Questionnaire Comments

1. How realistic was the simulation?

Controller Responses

S01 - Presentation on display, good. Aircraft/type vs. performance capabilities, needs editing.

S02 - The only problem that really stood out was the aircraft performance characteristics (i.e., speeds) were not very true.

S03 - Everything was realistic with the exception that speed inquiries were not accurate to ground speed.

S04 - Other than the aircraft mach#, pretty well.

S05 - Blank.

S06 - Outside of performance characteristics of aircraft being wrong, the simulation was very realistic.

S07 - In certain instances very realistic but on the average 1-5... "4".

S08 - Above average including the mistake by the sim-pilots because real pilots make mistakes also. You need to add radar clutter for more realism.

S09 - Somewhat. Apparently a good deal of effort was put into the design of the airspace, and the traffic scenario.

S10 - Somewhat, there were certainly parts that were not realistic (e.g., mach number correlation to altitude), but all in all the basic concept has been captured fairly well.

S11 - A) Mach #'s unrealistic. B) Air carriers do not file into VRB/SUA. C) MD88's can't make FL390. D) One problem had Miami Center landing Keystone 3 aircraft at FL330 with no plan.

S12 - Moderately - speeds, mach #'s, A/C performance (i.e., MD80 @ FL390) were not in line with real world. On F75/76 problems too many VRB descents (down arrow indicated).

S13 - It was good except for the speeds. Once we got used to the speeds it is all relative. Type A/C versus performance was unrealistic.

S14 - Fairly realistic. When asking for an aircraft on a heading the sims did not put them on the headings but gave us the control for turns instead.

S15 - Very realistic.

S16 - Moderately realistic!

S17 - I believe the simulation was not very realistic in the fact that it did not include aircraft performance characteristics.

S18 - Very close.

2. How representative was the generic sector of a typical en route environment?

Controller Responses

S01 - Good overall.

S02 - Very helpful, however after running the first generic problem it wasn't necessary to use it.

S03 - It was very close to sectors at the center.

S04 - The airspace was very representative, the ability to get information was hindered (i.e., range bearing, track heading, read out button).

S05 - Blank.

S06 - Shelf in CHARLIE CENTER was somewhat out of place.

S07 - The sector overall was relatively close to typical without clutter of ?primary? targets.

S08 - Fairly representative. Although I would add aircraft SE-NW flying fix radial distances. The letter of agreement has some problems:

1. J75 is shown as a one-way airway, although the letter discussed it as if it were two way.
2. Add aircraft as in 1A (1)(a)&(b).
3. P.O. PROC. with BRAVO sector should be over flights only developmentals should do their own POs for descending A/C - also, you need more noise.

S09 - Very. A good representation.

S10 - Good. It seemed to have a descent mix of overflight and departure/arrival traffic, we could have used a bit more traffic situations to work arrival/departures around. Also more of a variety of aircraft types (BE10's, PA46).

S11 - Good mix of traffic with departures and arrivals. Very realistic.

S12 - Very much. Good mix of routes/options/warning area shelves.

S13 - Good. Very straight forward.

S14 - Somewhat realistic. Letters of agreement were not as complicated.

S15 - Somewhat - traffic and sector bit too simplistic.

S16 - Fairly common.

S17 - A typical environment for an enroute controller is working a variety of sectors. Some are 0-230, some are 240-600, and others are stratified at various other levels. The generic sector is 240-600, which presents only one scenario.

S18 - Real to life.

3. How helpful was the training booklet in learning the generic sector?

Controller Responses

S01 - Good book, helpful.

S02 - Very helpful, however after running the first generic problem it wasn't necessary to use it.

S03 - It was somewhat helpful. There was a couple of contradictions in the LEAs.

S04 - Very, although we did expect north bound traffic on J75 according to the LOA.

S05 - Blank.

S06 - Map was helpful.

S07 - The training booklet was helpful but I basically learned the traffic flow and procedures during the problems.

S08 - See above - #2's response.

S09 - Not very. Some of the material was outdated.

S10 - The one forwarded to ZJX was a bit out of date, but still gave me an idea of the situations to expect. Any discrepancies were cleared up at day one's briefing.

S11 - Not very helpful. I learned the LOAs and other pertinent information when I got here.

S12 - A/S so easy to learn didn't need, but short glance.

S13 - I didn't have an opportunity until the last minute to review.

S14 - Very.

S15 - Little.

S16 - Somewhat helpful.

S17 - Didn't receive it.

S18 - N/A.

4. Was hands-on training adequate on the day 1 and day 2 session?

Controller Responses

S01 - Yes.

S02 - Yes.

S03 - Yes.

S04 - Yes, no problem.

S05 - Blank.

S06 - Yes.

S07 - Yes.

S08 - Yes, I look forward to the replacement PVD's. One idea, add function so when military airspace are hot, they can change colors.

S09 - Yes. Very sufficient.

S10 - Yes. Day 1 and 2 was enough to bring any FPL up to speed on the equipment and generic Hi? sector.

S11 - Yes, it helped me learn sectors and frequencies of generic center.

S12 - Yes, Also the problems repeated situations which made the situations redundant.

S13 - Yes.

S14 - Yes.

S15 - Yes.

S16 - Yes.

S17 - Yes.

S18 - Yes.



5. How could the generic sector be improved?

Controller Responses

S01 - 1) Turn off auto-data block positioner. 2) On departures, instead of showing FL180 as assigned, show requested altitude while aircraft is in climb from low altitude sector. 3) Don't auto center the track ball, leave it where it is.

S02 - If you could add a little more complexity (i.e., crossing traffic), even possibly add another airport at the north or south of the sector.

S03 - All traffic conflicts occurred in a couple of spots. You need to "mix it up", so that they don't expect the same problem over and over.

S04 - More crossing traffic, already at altitude not just climbers off MID.

S05 - Blank.

S06 - Add a restricted area to be avoided.

S07 - Aircraft speeds more realistic to types 1 or 2 altitude changes en route.

S08 - Speeds, if this cannot be corrected, teach the remote to adjust, this is critical. As developmentals will have a skewed perception of speeds and speed control that will be hard to correct.

S09 - Don't know yet. If I think of something you will be the first to hear of it!

S10 - (See #2 for some suggestions). Although it would be hard to truly capture the "REAL" thing in a simulation. Possibly adding control room noise, other controller/sector requests, clearance request changes to flight plans etc. would make it a more believable simulation.

S11 - A) Fix speeds. B) Possibly dual departures that we need to sequence. C) More airways.

S12 - More crossing airways and head-on traffic. Remove auto-point on shelf. Specify who (sector) receives H/O on UTN/DTN descents (down arrow indicated).

S13 - A) Possible wind conditions/WX inclusions. B) Have extended vector lines available to identify possible traffic conflicts.

S14 - ?.

S15 - Destination identifiers could be used in datablock. On screen "qak oak"?? was clumsy to use.

S16 - Design more difficult problems, more complexity!

S17 - Include aircraft performance characteristics. Build another sector 0-230 that incorporates sequencing departures out of uncontrolled airports with enroute traffic. Add VFR pop ups and IFR air files.

S18 - MOAs better airport ID's.

6. Did the ATWIT device interfere with controlling traffic on either sector?

Controller Responses

S01 - No.

S02 - No.

S03 - No.

S04 - No.

S05 - Blank.

S06 - No.

S07 - No.

S08 - No, we are used to distractions and are bored without them.

S09 - No.

S10 - Hardly any, once you got accustomed to it's frequency and requested data.

S11 - No.

S12 - No.

S13 - Blank.

S14 - No.

S15 - No.

S16 - No.

S17 - No.

S18 - Not really.

7. How well did the pseudo-pilots respond to your clearances in terms of traffic movement and call backs?

Controller Responses

S01 - Overall good. Sometimes, on controller to controller actions they weren't sure how to respond or what action they should take.

S02 - For the most part - very well.

S03 - The sim-pilots did an excellent job and sounded very much like the real thing!

S04 - The sim-pilots did a very good job, we even got bad read backs, keeping too true to life with real pilots.

S05 - Blank.

S06 - Extremely well, better than real pilots.

S07 - Good job.

S08 - As well if not better than real pilots and facilities, even mistakes (turns, readbacks, add realism.

S09 - Very well. A professional attitude and attitude were exhibited.

S10 - Almost without error. Very well done.

S11 - Good.

S12 - Very well except 1 problem which was later attributed to computer sim. lag.

S13 - My pilot's phraseology was terrible. He needs to review point out procedures.

S14 - Good, except that when asking the transferring controller to put the aircraft on a heading, they just gave us control instead of putting aircraft on the heading.

S15 - Good.

S16 - Very well.

S17 - Excellent.

S18 - Good response, but some of aircraft movements unrealistic.

8. Do you have any other comments about your experiences during the simulation?

Controller Responses

S01 - Thanks for allowing me to participate. Enjoyable and educational.

S02 - I enjoyed being a part of the experiment and would gladly return in the future if my services are ever needed again.

S03 - In general a good simulation, much better than I've experienced during the past training.

S04 - Very enjoyable, everyone was helpful.

S05 - Blank.

S06 - Don't like data blocks moving by themselves requested final altitude should already be in datablock, with T-altitude on climber.

S07 - I've had a good time with running problems, the atmosphere was very pleasant, if you need assistance once again I will be more than happy to come.

S08 - I think this will be a good screening/training device, if it is not used as a "Gothcha" program but used as designed.

S09 - No.

S10 - Best of luck in your research!

S11 - Blank.

S12 - A good learning tool that removes complexity of A/S knowledge. Excellent screen tool that already surpasses D1/D2.

S13 - Excellent. Staff and Instructors were extremely helpful. Again Thanks.

S14 - Everyone was knowledgeable and helpful. It made for a good experience.

S15 - No.

S16 - Most enjoyable.

S17 - It was wonderful. However the keyboard I was working with kept sticking and at times it became extremely annoying.

S18 - No.